A Framework for Data Loss Prevention using Document Semantic Signature

by

Hanan Alhindi
B.Sc. of Information Technology, King Saud University, 2009
M.Sc. of Computer Engineering, King Saud University, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

A Framework for Data Loss Prevention using Document Semantic Signature

by

Hanan Alhindi
B.Sc. of Information Technology, King Saud University, 2009
M.Sc. of Computer Engineering, King Saud University, 2013

Supervisory Committee

Dr. Issa Traore, Supervisor
(Department of Electrical and Computer Engineering, University of Victoria)

Dr. Kin Fun Li, Departmental Member
(Department of Electrical and Computer Engineering, University of Victoria)

Dr. Venkatesh Srinivasan, Outside Member
(Department of Computer Science, University of Victoria)

Supervisory Committee

Dr. Issa Traore, Supervisor
(Department of Electrical and Computer Engineering, University of Victoria)

Dr. Kin Fun Li, Member
(Department of Electrical and Computer Engineering, University of Victoria)

Dr. Venkatesh Srinivasan, Outside Member
(Department of Computer Science, University of Victoria)

Abstract

The theft and exfiltration of sensitive data (e.g., state secrets, trade secrets, company records, etc.) represent one of the most damaging threats that can be carried out by malicious insiders against institutions and organizations because this could seriously diminish the confidentiality, integrity, and availability of the organization's data. Data protection and insider threat detection and prevention are significant steps for any organization to enhance its internal security. In the last decade, data loss prevention (DLP) has emerged as one of the key mechanisms currently used by organizations to detect and block unauthorized data transfer from the organization perimeter. However, existing DLP approaches face several practical challenges, such as their relatively low accuracy that in turn affects their prevention capability. Also, current DLP approaches are ineffective in handling unstructured data or searching and comparing content semantically when confronted with evasion tactics where sensitive content is rewritten without

changing its semantic. In the current dissertation, we present a new DLP model that tracks sensitive data using a summarized version of the content semantic called document semantic signature (DSS). The DSS can be updated dynamically as the protected content change and it is resilient against evasion tactics, such as content rewriting. We use domain specific ontologies to capture content semantics and track conceptual similarity and relevancy using adequate metrics to identify data leak from sensitive documents. The evaluation of the DSS model on two public datasets of different domain of interests achieved very encouraging results in terms of detection effectiveness.

Table of Content

List of Figures

List of Tables

List of Abbreviations

BM Boyer Moore algorithm
CBSD Component-based Software Development
CERT/CC The CERT Coordination Center of Carnegie Mellon University
CF Concept Vector File
CT Concept Tree
DCT Document Concept Tree
DL Ontology Description Logics
DLP Data Loss Prevention
DSS Document Semantic Signature
EIC European Investigative Collaborations
FDR False Discovery Rate
FIBO Financial Business Ontology
FNR False Negative Rate
FPR False Positive Rate
IDS Intrusion Detection Systems
KB Knowledge Base
KDE Kernel Density Estimation
LSA Latent Semantic Analysis
NIDS Network-based Intrusion Detection System
NTAC National Threat Assessment Center
OWL Ontology Web Language
RDF Resource Description Framework
RNCVM Relevancy Nodes based Concept Vector Model
SEAM Semi-Automated Ontology Management
SIDD Sensitive Information Dissemination Detection
SVM Support Vector Machines
SW Smith Waterman algorithm
TF-IDF Term Frequency - Inverse Document Frequency

ACKNOWLEDGEMENT

First and foremost, I would like to thank my God Almighty, Allah, for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

There are no proper words to convey my deep and sincere gratitude and respect for my PhD supervisor Prof. Issa Traore for all his continuous support, guidance, and encouragement that he provided to me during my PhD. It was my pleasure working under his supervision and having a golden opportunity to learn from him. In addition, I would like to thank my supervisory committee for their valuable comments and encouragement.

My special thanks go to my wonderful father and mother, Dr. Waheed Alhindi and Mrs. Elham Abu Hassan, for their endless encouragement, unconditional support, continuous praying, and great love. Thanks go to my parents because they always believe in me as a successful strong woman. As well, thanks to them for their phone calls, from which I derived the strength to continue my work. Moreover, my honest thanks to my lovely grandmother Fatimah, who always prays for me to be a successful woman and obtain a PhD degree.

Also, I would like to thank Fahad and Reem Alossaimi, my husband and my daughter, who always believe in me and stand by me. Sincere thanks to Fahad for his continuous support and motivation that help me to follow my dreams. Thanks again Fahad for being a great husband,

kind father, best friend, loyal companion, and continuous supporter. My thanks to my lovely daughter Reem, who always tells me that she dreams to be like me and she wants to study at UVIC in the future.

My sincere gratitude and warm thanks to my brother Abdulrahman and my sisters Afnan, Reem, Raghad, and Sara for their emotional support, continuous encouragement, and frequent proud messages that reached to me over these years.

This journey and this big achievement would not have been possible without the support of my professor, parents, family, and friends. This research is supported by the King Saud University at Riyadh, Saudi Arabia.

DEDICATION

I dedicate this work to my parents, my husband, my daughter, and all my siblings.

Chapter 1

Introduction

1.1 Context

Malicious insiders are people with legitimate access to information, who abuse their privileges to damage or steal the organization's resources and assets [1][2]. Even though insider threat events are less widely publicized and relatively more infrequent than external attacks, they usually pose a much higher severity of risk for organizations when they do happen and can cause a tremendous amount of damage to an institution or a business. According to a survey conducted by CA Technologies on the state of insider threat in 2018 [3], 90% of surveyed organizations felt vulnerable to insider attacks and 53% of organizations indicated that they have been the target of inside attack during the year.

In previous attempts to understand insiders' threats, the US Secret Service National Threat Assessment Center (NTAC) and the CERT Coordination Center of Carnegie Mellon University (CERT/CC) conducted an extensive study analyzing insiders' threats from both behavioral and technical perspectives, across various sectors, including banking and finance, information and telecommunications, government, and critical infrastructures [5]. Some of the key findings of this study include the fact that most insider incidents involve little technical sophistication or complexity, and most of the incidents are driven by financial motives rather than the will to harm the organization.

There are two categories of insider detection approaches: behavioral and technological. Behavioral approaches use psychological profiles of perpetrators in pre-employment screening as

well as in monitoring the activity of technology specialists [6]. Technological measures consist of hacker monitoring devices or software for identifying a violation of internal security policy or abnormal use of internal resources. The main limitation of existing technological approaches is that insider behavior does not necessarily look abnormal and as such may not be detectable by traditional hacker monitoring devices.

Existing malicious insider detection systems consist of an apparatus of disparate security tools ranging from traditional intrusion detection systems (IDS), security events logging and auditing, to data leak prevention systems. However, traditional security devices are not well equipped to find sophisticated malicious insider attacks. For instance, NIDSs (Network-based IDSs), as we know them, are monitoring network traffic for known patterns of malicious behavior. They are completely useless in scenarios where there is no attack and no violation of communication policies. The same is true for firewalls, host-based IDSs, antivirus, and so on. All these devices are good at detecting malicious traffic or behavior. Insiders are authorized personnel, and as such their behaviors have all the hallmark of normality.

A common form of insider threat that affects a broad range of organizations, is an unauthorized data leak. Data Loss Prevention (DLP) is one of the most popular security controls used by organizations to fight against insider threat and prevent unauthorized data transfer [[1],[2],[3],[4]]. DLP consists of a security mechanism and/or strategy to prevent the illicit transfer by end-users of sensitive content outside the organization network. The scope of the current dissertation is developing a new DLP model to protect against illicit data transfer.

1.2 Research Problem

DLP consists of monitoring data transfer by end users to ensure that sensitive information is not sent outside the organization network. Existing DLP systems provide insider detection and prevention capability by implementing organizational, business and regulatory policies under the form of batteries of pre-defined or customizable rules. The monitored data is matched against the rules, and the decision is made whether there is potential for data leakage or not.

Existing DLP schemes track data leak by checking file names, data formats or specific keywords. DLP typically deals with two broad kinds of data: structured and unstructured. Structured data are data that fit predefined formats, such as social security numbers and credit card numbers. Unstructured data are data that do not obey any formatting restrictions or involve heterogeneous formats (e.g., media files, blog posts, emails, source code, etc.). Most existing DLP schemes focus only on tracking and matching structured data, which is commonly implemented using a set of rules and regular expression matching [1]. Unstructured data matching, which is more challenging, is performed by existing systems by computing and storing one-way hashes for protected content, and then tracking possible leak by identifying similar content in other documents. However, the shortcoming of using one-way hash for unstructured data matching is that this approach works only if an exact copy of the data is transferred; it is not effective in detecting situations where an altered, reworded (e.g. using synonymous or code words) or summarized version of the original data is leaked. Furthermore, detection based on regular expression matching can be evaded easily, as a malicious insider can skillfully remove from the data all problematic keywords, or even rewrite the content in a different language.

The shortcoming of such fingerprinting approach is that it is effective only if an exact copy of the data is transferred. Using a fingerprint based on cryptographic hashes will not detect situations where an altered or summarized version of the original data is leaked. Furthermore, detection based on regular expression matching can be evaded easily. A skilled malicious

insider, can clean the data by removing all problematic keywords, or even use a different language to rewrite the content while keeping the same meaning. As a result, existing DLP systems tend to generate a high level of false positives. Due to the high false alarm rates, most systems focus only on the detection and offer very limited prevention capability (i.e. avoid blocking suspected sensitive content being leaked).

The objectives of the current thesis is to address the aforementioned shortcomings by proposing a new signature scheme for unstructured data based on the data semantics, rather than the hash. This allows matching effectively altered or partially relevant/similar content against original content classified as sensitive or critical.

1.3 Proposed Approach

Our proposed DLP approach relies on a new document content fingerprinting scheme, termed document semantic signature (DSS). This is derived by extracting and summarizing the semantic or meaning of the knowledge contained by a file or other containers (e.g. email, repositories). The DSS is updated dynamically as the knowledge changes. Existing data loss prevention schemes monitor the file name or specific data formats or keywords contained in the file. These approaches fail to detect data transfer where the original information is altered, rewritten or reworded by using synonyms or code words. For instance, such systems will fail to detect a situation where an insider reads classified information (that she is allowed to read), and transcribes a summary of such information in a new container (file, email) using different lexicon or terminologies. In contrast, our proposed approach allows tracking malicious data transfers or exfiltration by monitoring the knowledge semantic (i.e. DSS) rather than the container (i.e. physical or logical file) or selected keywords.

Our content monitoring scheme uses domain ontologies to capture and encode the semantic of the knowledge or information being protected. An ontology is a formal representation of a set of concepts and the relations between these concepts in a domain of knowledge. Ontologies are used to reason about the instances and entities within the domain and to describe the domain. These provide the semantic (i.e. meaning) representation of the concepts and allow inferring the underlying relationships. We use domain-specific ontologies for the different kinds of knowledge being protected by a given organization (e.g. defense, healthcare, finance). There are a few existing insider-related ontologies in the literature, such as the *Insider Threat Indicator Ontology*, developed by CERT/CC [6]. However, these ontologies describe domains and concepts related to the creation and operation of insider prevention tools, policies, and models. We are interested in describing the actual knowledge or data that is being protected, and as such, we use ontologies specific to the corresponding knowledge domains.

The DSS is derived by extracting a summarized representation of the semantics model for a given file or content. The system will monitor newly generated contents (new email being written, or new file creation) and tracks malicious data transfer between user accounts or data exfiltration between an insider's account and output channels (e.g., emails, printers, online repositories) by comparing the DSS of the transferred data against the DSS of the critical documents.

The system will maintain a pool of DSS for files containing classified information. The classification of newly generated content will be determined automatically by analyzing the content semantics and comparing the corresponding DSS against the pool of classified documents' DSS. The comparison of a DSS against a pool of DSS is done by using semantic similarity and relevancy metrics.

Experimental validation of the system will be conducted by calculating the following performance metrics:

- detection rate (the ability of the system to detect insiders),
- false positive rate (the error rate in flagging an insider activity as malicious while such activity is not malicious).

The goal will be to achieve the highest detection rate while minimizing the false positive rate.

1.4 Thesis Contributions

The main contributions of this dissertation are as follows:

1. Development of a new approach for tracking and detecting sensitive data leakage that consists of a summarized representation of the content semantic called the document semantic signature (DSS).
2. The ability, through the DSS, to detect data leaks for documents with complex content, in terms of the diversity of the knowledge.
3. Development of a new approach of data loss prevention (DLP) that tracks and matches unstructured data based on the content semantics. This allows matching effectively altered or partially relevant/similar content against original content classified as sensitive or critical.
4. The ability, through the DLP, to detect attempts at evading detection, such as rewriting or modifying the content while keeping the meaning of the information unchanged.
5. A new dataset for DLP research has been collected and organized that combines public data under the form of Football news articles and private Football leaked data.

The proposed model has been evaluated experimentally using two different datasets and two different ontologies of different knowledge, as follows:

- The business domain of interest.
- The sport domain of interest.

The evaluation results show that the proposed approach achieves high detection effectiveness in terms of accuracy.

Contributions 1 and 3 have been published in the 2nd International Conference on Wireless, Intelligent, and Distributed Environment for Communication (WIDECOM 2019), and received the best paper award [8].

Contribution 4 has been published in a journal paper titled Internet of Things: Engineering Cyber Physical Human Systems, Elsevier, 2019 [9].

Contributions 3 and 5 will be submitted to a conference [60].

1.4.1 List of publications

1. Alhindi H., Traore I., and I. Woungang, “Preventing Data Leak through Semantic Analysis”, Journal of Internet of Things, Elsevier (In Press, Accepted 25 June 2019).
2. Hanan Alhindi, Issa Traore, and Isaac Woungang, “Data Loss Prevention Using Document Semantic Signature”, International Conference on Wireless, Intelligent, and Distributed Environment for Communication (WIDECOM 2019), Milan, Italy, Feb 11-13, 2019, Springer, Best paper award.
3. Hanan Alhindi, Issa Traore, and Isaac Woungang, “Preventing Data Loss by Harnessing semantic similarity and relevance”, to be submitted to International Conference on

Wireless, Intelligent, and Distributed Environment for Communication (WIDECOM 2020), to be held May 06-08, 2020, Toronto, Canada.

1.5 Thesis Outline

The remaining chapters of the thesis are organized as follows:

Chapter 2 provides a broad background on data loss prevention including data loss causes, data loss prevention methods categorization, and a comparison between different data loss prevention methods. Also, it gives an overview of the literature and introduces related works on insider threat detection, data loss and leakage prevention, and ontological search.

Chapter 3 presents and describes in details our proposed DLP model including the semantic similarity and relevancy metrics that have been applied.

Chapter 4 describes in details the experimental evaluation of the proposed model on the different datasets and using different semantic metrics, and discusses obtained results. In addition, the experimental evaluations of baseline DLP models are presented and compared with our proposed model.

Chapter 5 concludes the thesis by summarizing the overall contributions of our research and outlining our future work.

Chapter 2

Background and Related Work

The field of data leak prevention approaches and technologies is relatively new, and as a result the relevant literature is limited. In this chapter, we give a broad background on data leak prevention concepts and techniques and summarize and discuss the research literature on data leak prevention.

2.1 Background on Data Leak Prevention

The increasing online presence of organizations and the growing cyber activities that this entails has led to a steady growth of their data [10]. Having massive amounts of critical data related to the organization, employees, and customers require high internal and external security to exchange data smoothly and safely. As a result of the aforementioned, organizations need to provide easy data communication for their users while protecting and preventing sensitive data from any breach. The breach could be in a form of data loss, which threatens organizations' security, competitiveness, and credibility.

Numerous data loss incidents have had negative impacts on the corresponding organizations. For example, the American multinational corporation Morgan Stanley faced an incident of data loss caused by one of its financial advisors called Galen Marsh, who stole critical accounts information of the company's wealth management clients [11]. As well, before a Walt Disney CEO announced the company's quarterly earnings to the public, he sent an email

about that to a reporter by accident in 2000 [12]. Also, a sensitive email about Prince Charles' visit to Poland was sent by mistake to a wrong email address in 2002. As a result of that situation, Prince Charles' plan was disclosed to newspapers before he announced that [12]. Moreover, Marriot International hotel group had experienced a massive data breach since 2004 by unauthorized access to its Starwood database. They reported that around 500 million customers' information were copied and disclosed [13]. All these incidents emphasize that protecting and preventing organizations from data loss is a significant step in terms of security and privacy.

2.1.1 Data Loss Prevention

DLP is a technique for detecting and preventing unauthorized disclosure, breaches, or exfiltration of the company's confidential information. DLP helps to protect sensitive data by preventing all types of data loss including intentional and unintentional data loss that could affect the confidentiality, integrity, or availability of the organization's sensitive information [14]. Some DLP techniques are used for detection, prevention, or both of the above. DLP techniques often consist of three main elements as shown in Figure 2.1 [10]:

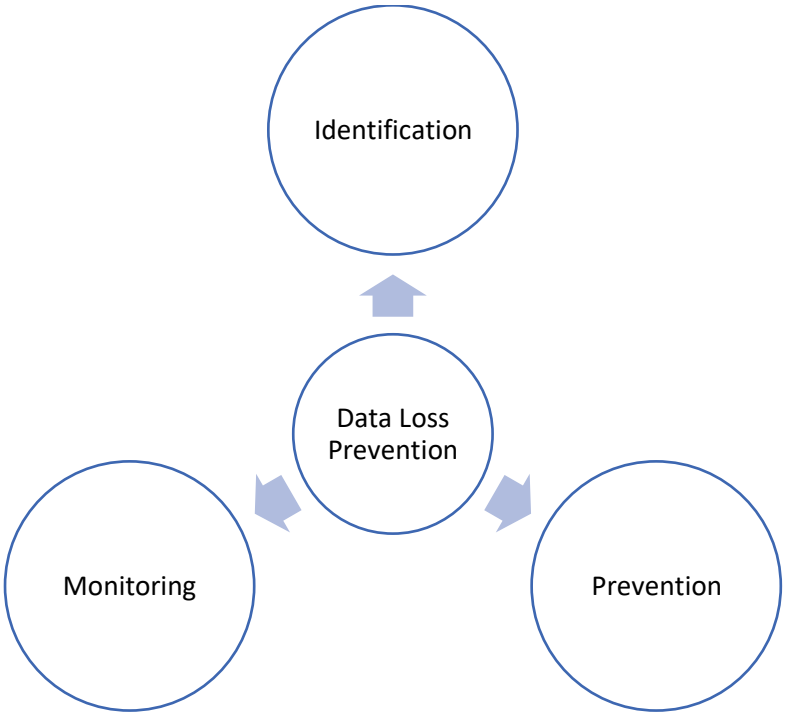


Figure 2.1 The three main elements of DLP technique[10]

1- **Identification** is a detection method of organizations' critical information based on predefined policies. Whether the DLP technique is just for detection, prevention or both of the above, this method usually consists of an analysis task that is based on either data context, content, or a combination of the above.

2- **Monitoring** is a tracking and flagging method of organizations' critical information that should not leave organizations' networks.

3- **Prevention** is a method of taking actions with flagged critical data that is derived after applying identification and monitoring methods. Different prevention methods are shown in Figure 2.2 including allowing users to access files or send emails, blocking the access to suspicious files or emails, labeling users and withdrawing some privileges, reporting the users' activities to the network administration or manager, and warning users about their current suspicious activities. These prevention methods could be applied either

individually or in combination when the organization's security policies are violated based on the organization's regulatory requirements.

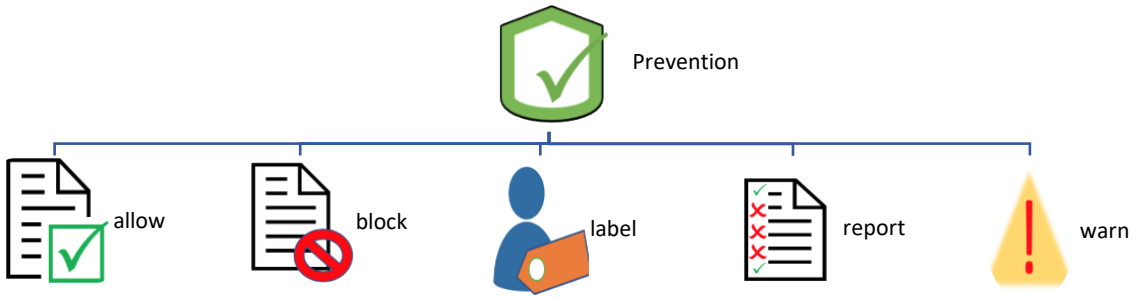


Figure 2.2 Several prevention methods of DLP

The organization's confidential information that should be protected and secured can exist in any state based on the data lifecycle as follows [10] [15][16]:

1- *Data at rest* is the data that is stored on company devices and systems such as databases, servers, file systems, computers, and distributed desktops.

2- *Data in use* is the data that is used and being processed by company users on endpoint devices such as a file being copied to a USB device, a file being accessed in a computer, or an email being written but not yet sent.

3- *Data in motion* is the data that is moved through the company network to the outside world by e-mails, instant messages, web traffic, or other communication methods.

The DLP technique helps to monitor, identify, and protect critical information from a loss or leak outside the corporate network. Data loss may occur through regular storage, usage or transmission data channels. In this situation, it is significant to identify the most popular data loss channels related to data states to provide the desired security level without impeding information flow through those channels. The common data loss channels for data at rest and data in use are through hard copies of documents, removable storage devices, and portable devices. Moreover,

data in motion could be leaked while sharing files and information or web services via emails, webmail, instant messages, social media, cloud, and portable devices. Figure 2.3 shows some examples of common data loss channels.

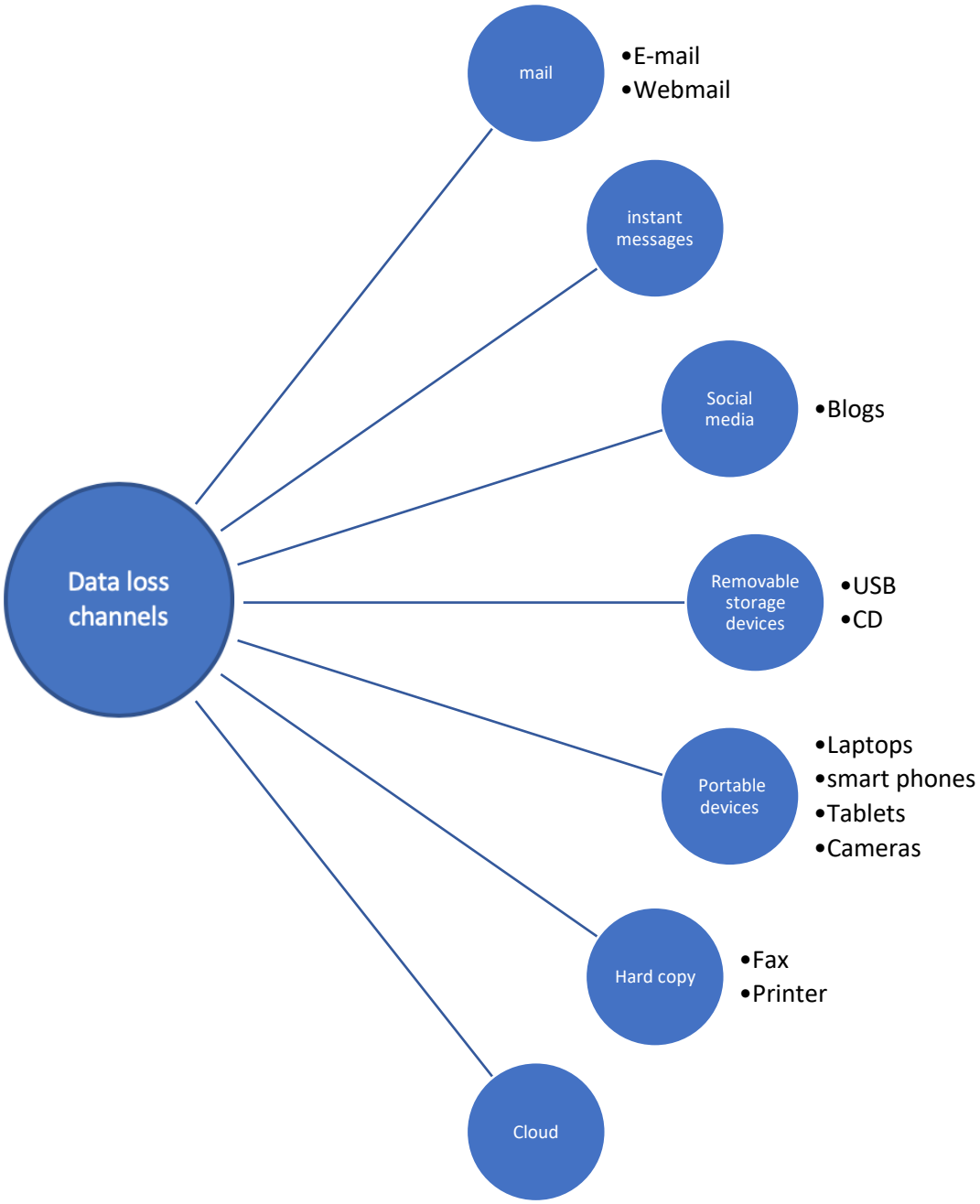


Figure 2.3 Common data loss channels

2.1.2 Data Loss Causes

There are two main causes of data loss in an organization based on human factors and their location according to an organization:

1- ***External cause:*** One of the main external causes of data loss is remote attackers. *Remote attackers* are the outsiders, who do not have access to organizations' data but they can penetrate the system by malware, injected codes, or social engineering attacks to illicitly access sensitive data and cause data loss to an organization [17].

2- ***Internal cause:*** One of the main internal causes of data loss is insider threats. *Insider threats* are authorized employees, who can intentionally abuse their privileges and maliciously access and transfer sensitive data out of organizations' network boundaries. Those insiders can transfer critical information to outside organizations' networks by either of two ways:

   a. *electronically* by sending information via web or e-mails, or

   b. *physically* by sending information via data storage devices such as USBs and hard disks.

Malicious Insider threats could cause data loss into two forms as follows [16] [17]:

1- *Damaging* critical information, which compromises the availability or the integrity of data by either hiding or corrupting the original correct copy of data.

2- *Leaking* critical information, which compromises data confidentiality by unauthorized disclosure.

2.1.3 Data Loss Categorization

Data loss incidents could happen in any organization either intentionally or unintentionally. To illustrate, insider threats such as disgruntled employees could cause intentional data loss by stealing sensitive organization's data, credit card records, or social security numbers and use them maliciously to destroy the organization's data, system, or finance. On the other hand, unintentional data loss could happen accidentally by natural disaster and fire, which could harm the system or data. In addition to that, employees could by mistake attach a critical document to an email, send an email to a wrong address, publish or post organization's private information to the public accidentally or negligently. Figure 2.4 shows the categorization of data loss.

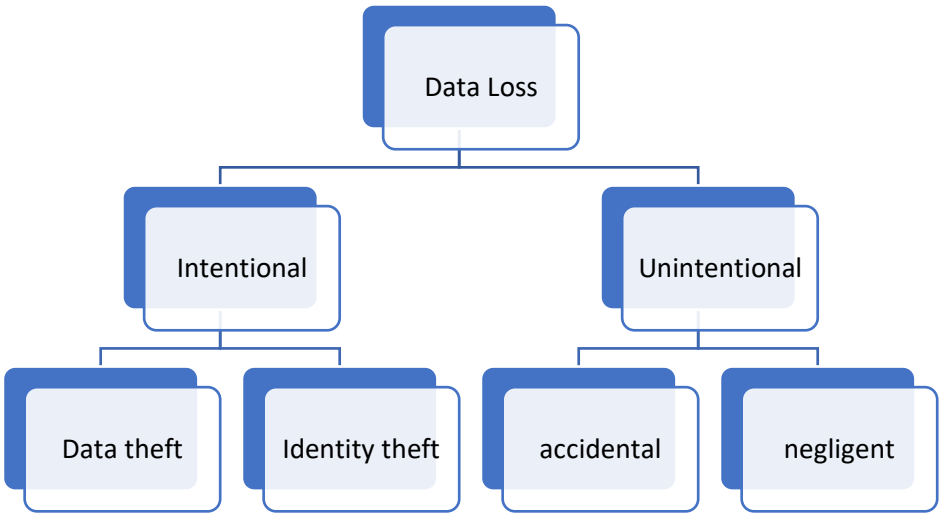


Figure 2.4 Data loss categorization

2.1.4 Data Loss Prevention Approaches

DLP approaches are categorized based on analyzing data context, content, or a combination of them. We can consider the content like a letter while the context as an envelope.

2.1.4.1 Context Analysis

The DLP techniques, which are based on context analysis, focus on the contextual attributes that are linked to the monitored data, such as source, destination, time, size and format [18]. Examples of context analysis DLP techniques include social and behavior analysis, data identification, and data mining and text clustering.

2.1.4.2 Content Analysis

The DLP techniques, which are based on content analysis, focus on and inspect the content of monitored data whether it is a document or an email to trigger policy violations in an organization. In this case, those DLP techniques are mainly based on three content analysis types, which are data fingerprinting, regular expression, and statistical analysis [19].

Table 2.1 presents and compares several DLP techniques by highlighting the underlying goals, and advantages and disadvantages [18] [20]. Figure 2.5 shows the categorization of data loss prevention approaches based on content and context analysis. Some of these approaches are based on either context, content, or both of them. Also, we classified DLP techniques based on the data that they deal with including data in-use, in-transit, and at-rest. We represented each DLP technique by its id that is shown in Table 2.1. For example, statistical analysis (Term weighting) approach with id 11 is based on content analysis, which deals with data-in-transit while policy and access rights approach with id 1 is based on context analysis and deals with data-in-use. When policy and access rights approach is combined with virtualization and isolation approach, they analyze data in use and data in transit contextually. Also, data mining and text clustering technique with id 7 could analyze the content and the context of data in

transit. Furthermore, Figure 2.6 classifies DLP techniques into detection, prevention, or detection and prevention techniques representing each technique by its id as shown in Table 2.1 [18].

Table 2.1 Comparison of DLP techniques.

DLP Id	DLP techniques	Technique goal	Advantages	Disadvantages
1	Policy and access rights [61][62]	Prevention	• Most common preventive technique for organizations. • Simple to implement and easy to control. • Uses contextual analysis. • One of the robust techniques to prevent sensitive data leak before it happens.	• Only a prevention technique. • Should be combined with a detection technique to detect and prevent sensitive data leak. • When access control and policy is being used, the technique is influenced.
2	Virtualization and Isolation [63][64]	Prevention	• When a user accesses sensitive information, a trusted virtual environment is created to isolate the user's activities and grant him privileges for certain actions.	• Only a prevention technique. • Should be combined with a detection technique to detect and prevent sensitive data leak. • Is expensive to implement and maintain.
3	Cryptographic approach [65]	Prevention	• Works by generating cipher text from a plain text of sensitive content. • Can prevent users from reading the plain text.	• No guarantee of security for the ciphertext.
4	Quantifying and limiting [66][67][68]	Prevention/ Detection	• Able to detect and prevent leaked data. • Administrator utilizes quantifying methods to build and organize critical data. • Interested in leaking channels	• Could interrupt data flow of some channels. • Does not guarantee to fully block leaking channel. • It has limited detection of leaked data via hidden channels.
5	Social and behavior analysis [69]	Detection	• It is considered a proactive prevention technique. • Can detect malicious relationships between people, groups, and organizations. • Can predict and track human behavior.	• Generate a high false positives rate. • Require administrator's monitoring and involvement.

				• Require collecting users' profiles for comparison and detecting irregular activities.
6	Data identification [55] [70][71][72]	Detection	• When this technique uses fingerprinting analysis, it generates a low false positive rate. • Very robust to detect unaltered data. • Can use some powerful hashing technique to detect altered data.	• Prior knowledge of sensitive data is required. • Cannot detect highly altered data. • Cannot understand data semantics.
7	Data mining and text clustering [30][73][74]	Detection	• Associated with machine learning techniques • Able to detect sensitive content of unstructured data • It is also a data leak prediction technique. • Able to perform a complicated task.	• May generate a high number of false positives. • Involve a massive amount of computations. • Has limited scalability
8	Data fingerprinting (exact/partial matching) [55] [72]	Detection	(Exact data/file matching) • Uses database for storing exact data or hash values for files • Effective for detecting structured data from the database. • Effective for fingerprinting any type of file. • Produces a low false positive rate. • Uses contextual analysis (Partial file matching) • Able to detect sensitive content of unstructured data • Works either for the whole file or part of the file • Generates a low false positive rate. • Uses Content analysis	(Exact matching) • The performance is influenced by the large size of the database. • Unable to detect unstructured data. • Unable to detect modified files. (Partial file matching) • It is necessary to specify critical documents that need to be protected. • The performance is influenced by the massive volume of critical content that needs to be protected.
9	Regular expression (dictionary-	Detection	• The analysis can be done by comparing the content against specific rules. • It is effective for detecting structured data such as credit card numbers, social security numbers	• It generates a high number of false-positive rates. • It has limited capability to detect sensitive information within unstructured data.

	based match) [75]			
10	Statistical analysis (N-gram analysis) [76]	Prevention/ Detection	• Uses Content analysis • Able to detect sensitive content in unstructured data • May classify the importance of the content based on machine learning techniques.	• Involve a large amount of data • May generate a high amount of false-positives and false negatives.
11	Statistical analysis (Term weighting) [77]	Prevention/ Detection		

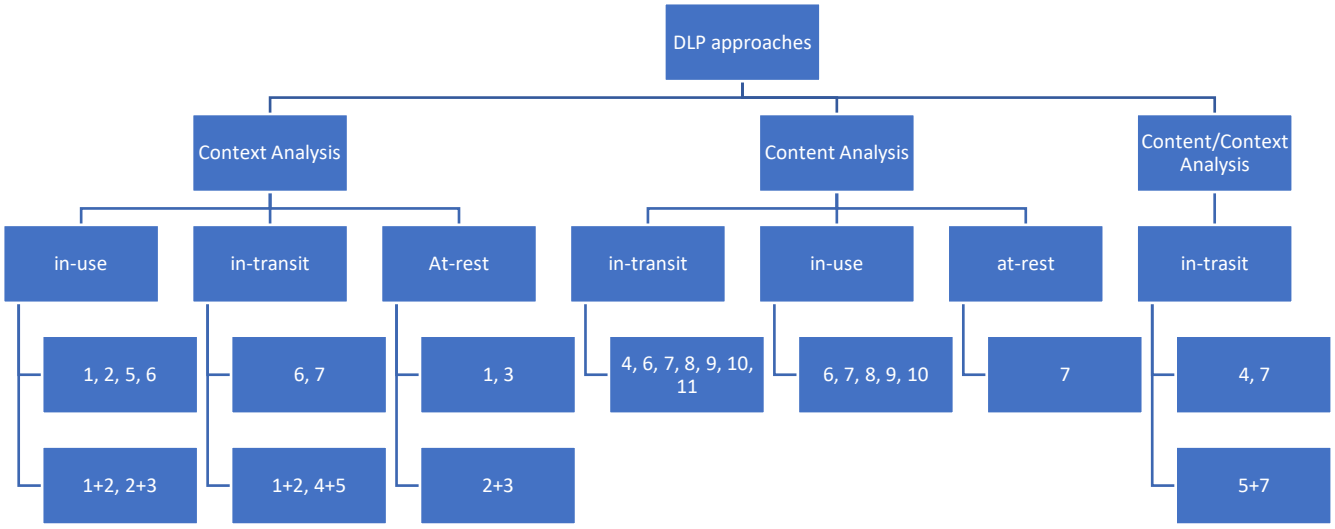


Figure 2.5 DLP categorization based on content/context analysis and data states; the numbers refer to the row ids in Table 2.1

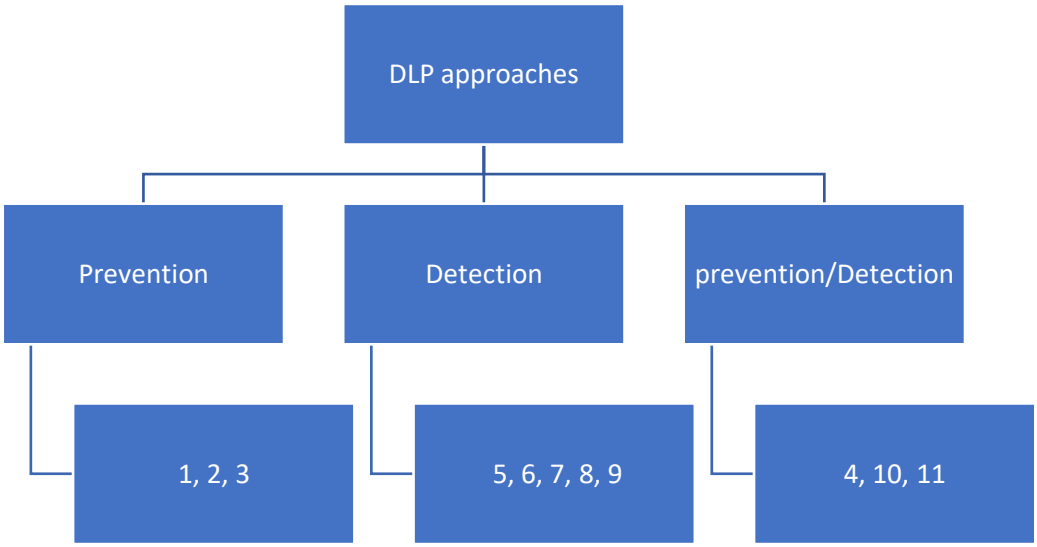


Figure 2.6 DLP classification based on detection/ prevention methods; the numbers refer to the row ids in Table 2.1

2.1.5 Existing DLP Solutions

Existing DLP solutions are available either as a full suite or a channel [10]. A full suite is a tool with full DLP solution, which has detection, prevention, and central management console components. This kind of DLP is mainly and widely used for preventing critical data loss in all data states. Also, full suite DLPs can focus on and deal with several network protocols including email, FTP, HTTP, and HTTPS. Fingerprinting technique is commonly used as a detection method for structured and unstructured data. In addition to a full suite, channel DLP is a tool with a limited DLP functionality along with other features in the product. Some channel DLP solutions could detect emails' sensitive content by using the pattern matching method.
There are several examples of DLP solutions currently available on the market such as CISCO, SonicWALL, Symantec, and McAfee [21].

2.1.6 Consequences of Data Loss

Lost data and service disruption require spending more time and efforts to repair, backup, and restore the original data and system. In addition to that, several consequences could happen including business functions disruption, financial losses, reputation and brand damage, and regulatory violation fines [18] [22]. The University of Texas reported that "94% of companies suffering from a catastrophic data loss do not survive – 43% never reopen and 51% close within two years" [23].

2.2 Related Works

In this subsection, we summarize related works on insider threat detection and prediction, data loss and leakage prevention, ontology-based search and information retrieval, and ontology management and semantic models. Sample notable works are reviewed and discussed under each category.

2.2.1 Insider Threat Detection and Prediction

Several insider threat prediction and detection models have been proposed in the literature. Most of them focus on analyzing user activity logs at the host level [24] [25] or at the network level [26].

Kandias et al. proposed a hybrid insider threat prediction model that combines real-time usage profiling and psychological profiling, along with user taxonomy [24]. Real-time usage profiling involves monitoring users' behaviors by analyzing system calls and extracting a behavioral pattern for each user. Psychological profiling is based on social learning theory, and involves measuring the user sophistication and predisposition, which are done by questionnaire, and user stress level, which is done by psychometric test. The proposed prediction model uses the extracted information from real-time usage, psychological profiling, along with user taxonomy as inputs to decide and identify possible malicious behavior. The decision-making algorithm can predict and score suspicious insiders based on three factors, which are motive, opportunity, and capability.

Udoeyop developed an insider detection approach by tracking threatening abnormal behavior through user activity monitoring [25]. In the proposed approach, user activity

information related to hardware, process, network, and file system, are collected. Various features are extracted from the data, such as processor usage, memory usage, hard drive usage, process threads, file system, network IP, and network port profiles. Normal behavioral profiles are constructed for each user using K-means clustering and Kernel Density Estimation (KDE) algorithms. In this approach, any new user behavior is collected, compared to user normal profile, and flagged as abnormal if there is a significant deviation.

Ragavan (2012) introduced an insider threat mitigation model that monitors and prevents malicious write operations on sensitive data by using log analysis and dependency graph [27]. The proposed log model stores each write operation that is done on any data item into log files, and then checks the number of changes on that data item according to the assigned threshold for each data item. If the number of changes on a data item by write operation exceeds the assigned threshold, then the system signals that threat and checks the validity of the write operation. As a result of non-sequential checking operation and its related delay, the author developed a dependency graph model that helps to save more time. The dependency graph consists of nodes corresponding to different data items in the database and edges representing dependency relationships between data items. In fact, this dependency graph is built after the transaction of write operation starts, which is more efficient to catch insiders. The author refers to the sensitive data items as *Critical Data Items* and the non-sensitive data items as *Regular Data Items*. Along with a dependency graph, the model assigns a particular threshold to each *Regular data item* in the database as a limit of the number of changes by the write operation. The *Critical data item* has more priority than *Regular data item* by checking the transaction immediately after any write operation. While if the number of write operations of *Regular data item* exceeds its threshold, the system signals the insider and checks the validity of write operation. As a result of a

dependency relationship between two nodes, any writing operation on a *Critical data item* or *Regular data item* that could affect other *Critical data item* should be secured and immediately checked by the system to validate the write operation. The limitation with this proposed model is that it was tested on a synthetic dataset, which could miss out important characteristics of real-world datasets.

Liu et al. (2009) developed a multilevel framework named SIDD (for Sensitive Information Dissemination Detection) system, which is mainly aimed to detect and prevent sensitive data leakage in a protected network [26]. The proposed detection system is placed at the egress point of the network to track outgoing traffic in order to filter transferred sensitive content. Network traffic features are analyzed to detect the existence of a covert channel. Some threats could use steganography techniques to hide the fact that communication has occurred. In this case, the system used steganalysis to detect these hidden channels.

Although the above approaches take a broader look at user activity (i.e. system calls, processes, network), in our proposed research we focus on leakage related to specific content, which is much narrower, but still very important, in scope.

2.2.2 Data Loss and Leakage Prevention

Raman et al. outlined the importance of the data leak prevention discipline by reviewing previous research, defining unsolved problem, introducing the challenges behind this problem, and motivating academic research to find a solution [28]. Raman et al. clarified that the goal behind data leak prevention is to detect and protect the resources, while intrusion detection goal is to detect the illegitimate users and protect the system from their activities. Liu and Kuhn

discussed data loss prevention challenges and defined several types of lost data including leaked, disappeared, or damaged data [29]. In addition, three data loss modes have been identified including data at rest, data at the endpoint, and data in motion, in order to find best practices and solutions capabilities to address the underlying problem.

Hart et al. (2011) proposed a DLP approach that relies on using a machine learning algorithm, specifically Support Vector Machines (SVM), to learn and automatically classify sensitive information, both structured and unstructured [30]. The monitored information is classified as either public or private. Models are trained using an initial set of public and private documents; the trained classifiers are later used to recognize sensitive (i.e. private) documents from non-sensitive ones. They introduced a new training technique, so-called *supplement and adjust*, that enables better discrimination between sensitive and non-sensitive data. The proposed approach was evaluated using 5 different datasets, yielding on average a false positive rate (FPR) of 0.46%, a false negative rate (FNR) of 1.6%, and a False Discovery Rate (FDR) of 0.47%. FDR is defined as the ratio between the number of false positive and the sum of the numbers of false positive and true positive. The authors claimed that they have developed the first publicly available corpora for DLP systems evaluation. However, when contacted, the authors were only able to point to the Enron email dataset, which was published elsewhere. They mentioned they could not share the private (or sensitive) subset of their evaluation corpora due to privacy and confidentiality reasons.

One of the key limitations of using machine learning for DLP is that for the model to be effective, enough representative samples of the sensitive data must be available to train the classifier, and the classifier may need to be retrained every time there is a significant change in the characteristics of the sensitive data. In contrast, our approach does not have such constraint

as it depends only on the semantics of the data, regardless of its amount and future changes. Our model dynamically updates the semantic model as the content of the sensitive data evolves.

Stamati-Koromina et al. proposed a data leak prevention model, which is able to detect sensitive leaked data via e-mail messages using steganography [31]. When the user sends an e-mail, the system scans, monitors, and logs the outgoing e-mail and its attachments. This model uses SMTP Proxy server along with another online tool to get attached e-mail's images and check if there is any embedded sensitive data inside these images. If the system detects a steganography payload in the attachment, the system prevents sensitive data leakage by directly marking the e-mail as sensitive, sending an alert to the administrator and terminating the e-mail transmission.

Canbay et al. developed a data leakage prevention system in Turkish language [32]. In the proposed approach, the model is trained initially by generating a list of sensitive words from sensitive documents by computing and analyzing the Term Frequency Inverse Document Frequency (TF-IDF) metric. Detection is then carried out by comparing a monitored document against the trained model, and is aimed at locating any modification on sensitive words including adding, deleting, and altering characters. The detection relies on using the Boyer Moore (BM) algorithm to search for explicit sensitive string, and the Smith Waterman (SW) algorithm to detect altered sensitive string. The proposed approach was evaluated on a dataset consisting of 180 documents covering different topics, yielding 100% recall and 98% accuracy. A key limitation of the proposed approach is the inability to detect modification where the semantic remains unchanged, e.g., by replacing a word by its synonymous. In contrast, our proposed approach can detect such form of data leakage as it relies on monitoring the content semantic.

In 2015, a novel semantic similarity detection approach for data leakage prevention was proposed by Du et al. [33]. This approach depended on the latent semantic analysis (LSA) and support vector machine (SVM) for sensitive semantic features' extraction to represent concepts. Also, removing stop words and stemming are applied and the relative frequency for each tested document is calculated and compared to a particular threshold to determine semantic similarities and prevent critical data leakage. Du et al. evaluated their model on five document sets and obtained performance rates ranging between 76.1% - 98.6% for TPR and 0.8% - 15.1% for FPR.

Similarly, in our approach, we implemented several steps such as removing stop words, stemming, and calculating the frequency to determine the most representative word in each document in the form of a set of concepts. However, Du et al. introduced the relative frequency by calculating the number of a specific term occurring in a tested document with respect to the number of words in that document. Instead, we introduced the frequency by calculating the number of matched concepts in a tested document with respect to a critical reference document. In addition, our proposed model utilized a lexicon of a specific domain of interest, which can provide more concepts' synonyms while extracting terms from tested documents and searching for concepts in an ontology. Moreover, Du et al. used a general dataset, which did not include leaked data and consisted of published research papers from Google scholar, IEEE, and ACM digital library, while we used two real-life datasets that contain critical leaked data.

Shapira et al. proposed an extended method for fingerprinting content to detect any data leakage [55]. The authors focused only on critical contents to generate fingerprints to produce less false alarms. Also, this approach is considered the first research in using k-skip-n-grams for text fingerprinting. In addition, this approach can detect data leakage in modified documents. The evaluation of the approach was conducted using two datasets consisting of Reuters news

articles and a subset of the PAN plagiarism corpus 2010. Under some scenarios, the obtained performance results were acceptable and the accuracy was high. However, there is a need in the approach to improve space efficiency while maintaining a high accuracy level. In addition, the evaluation lacked real-life leaked datasets; this would be required for more accurate results. Also, once the leaked data is detected, a proper prevention method should be incorporated to protect critical data.

Costante et al. introduced a hybrid framework for data loss prevention and detection, which includes five main steps: learning, prevention, detection, alert analysis and rule management [56]. Also, the framework aggregates both anomaly-based and signature-based components. Once users' activities are monitored, an anomaly-based component is used for detecting transaction with abnormal behaviors. Alerts for malicious transactions will be flagged and blocked by rule-based prevention technique. In the meantime, a signature-based component is used for generating an attack signature from malicious alert to prevent any upcoming similar transactions. Costante et al. evaluated the framework on two datasets, which are synthetic and real-life datasets obtained from GnuHealth and Oracle, respectively. The authors studied the relation between the number of attacks and time and they found that the framework achieved a quick response in triggering the prevention rules for malicious transactions. However, the accuracy of the approach was not studied; important performance measures, such as detection rate and false positive rate have not been provided.

Kaur et al. proposed a new data leakage prevention algorithm via gateway to protect transferred sensitive emails from any disclosure [78]. The proposed algorithm consists of several steps including creating a sensitive keyword list for each department, checking the email subject, content, and attachment, checking the sender's department, matching extracted keywords from

the email with the sensitive keyword list, and preventing the transfer of sensitive email if similarity is determined. The prevention methods that have been introduced are either blocking, encrypting, or quarantining email. The limitation with this method is that it works only with emails but not with other kinds of knowledge containers such as files, blog posts, and repositories. However, our proposed DLP approach supports different kinds of knowledge containers by monitoring and detecting sensitive data leakage from any newly generated content such as new emails being written and new files being created.

Ling et al. proposed a new method for network data leakage prevention by checking TCP data packets [79]. The proposed method stores in a server users' names along with the sensitive information that they are allowed to transfer. Then, the system retrieves the source user information and the sensitive content in any TCP packet being transferred via the network. After that, the system checks if the source user of a transferred packet is eligible to send the sensitive information in that TCP packet. Next, based on user policy the system will either allow or block transmitting the TCP packet. The limitation on this method is that it allows an authenticated user to transmit TCP packet, which carries a sensitive information; however, the system is not able to detect or prevent sensitive data leakage if an authenticated user discloses sensitive information through that transmission. In other words, the system is able to detect sensitive data leakage if an unauthorized user sends sensitive information via TCP packet.

2.2.3 Ontology-based Search and Information Retrieval

Our proposed approach uses ontology-based search capabilities to search for terms and their semantics. There is a rich body of literature on ontology-based search. However, the proposed

approaches cover information retrieval from a general perspective, without any particular focus on data loss prevention.

Vodithala and Pabboju proposed an ontology-based search approach that relies on searching and retrieving information based on a keyword and associated semantic keywords **Error! Reference source not found.**[34]. The proposed approach was used in Software Engineering in the context of Component-based software development (CBSD). CBSD aims to maximize reuse of software components, which are pieces of code, to save time and reduce development cost. In this approach, each software component is stored in the repository in a file along with several keywords that are related to the component and describe it. These keywords are supposed to be arranged in an ontology in a tree form for searching method. Once the user searches about a specific keyword in the ontology, the system extracts the concept and all its siblings' concepts as exact matches, while its children concepts are considered as approximate matches. Then, the system retrieves the components corresponding to all the retrieved concepts from the repository. This work inspired us in our proposed model in providing a lexicon of keywords (i.e. dictionary of concepts and their synonyms) when the system searches about a keyword on ontology to retrieve related semantic keywords.

Fernández et al. proposed an ontology-based information retrieval model, which involves indexing, querying, searching, and ranking phases, to enhance the semantic search in the web environment [35]. The enhanced information retrieval model uses the domain knowledge Base (KB) with SPARQL query to get a set of tuples, which are used to retrieve the documents that contain a keyword and its related semantic keywords from large document repositories. Our proposed model uses similar retrieval technique, but we do not use SPARQL, which turns out to be difficult to configure and not adequate for our purpose. Instead, we use an ontology-based

search approach that retrieves words and their related semantics by looking for these terms through two main ontology components, which are class and definition components.

2.2.4 Ontology Management and Semantic Models

Doing-Harris et al. (2015) developed an ontology management system named SEAM (for Semi-Automated Ontology Management) that provides information extraction from clinical and biomedical documents based on OWL ontology files [36]. SEAM uses natural language processing to extract terms and their relations. In the proposed approach, a TF-IDF (Term Frequency-Inverse Document Frequency) vector is generated for each N-gram, containing one entry for each document. An entry (i.e. TF-IDF value) in the vector for a given document is the occurrence frequency of the N-gram in the document divided by the average frequency across all documents.

In contrast, our proposed model extracts information from documents based on RDF ontology files and generates a collection of document vectors obtained by retrieving the depth of all terms with respect to a specific term in the ontology, and where each vector corresponds to a unique concept from a document. The document vectors are combined into a matrix that represents the document semantic signature.

Liu et al. proposed a semantic model, the *Relevancy Nodes based Concept Vector Model (RNCVM)*, where a concept vector is used to represent a particular concept node in a hierarchical structure [37]. In this case, the concept vector of a specific node is based on *local density of all relevancy nodes* in a taxonomy structure. In contrast, in our proposed model, we create the concept vector by defining the *depth of all relevancy nodes* of a specific node in the ontology. In

our proposed work and the paper by Liu et al., the similarity between two concept nodes can be measured by using their concept vectors. Relevancy Nodes based Concept Vector Model with WordNet achieved higher correlation value of 0.906 with human judgments when compared to several existing similarity measures. However, the model finds the similarity based on WordNet which sometimes gives several synonyms that are not related to the domain of interest. Because of that, in our proposed model we provide a lexicon for specific domain of interest in order to facilitate retrieving the right synonyms of a keyword.

Li et al. introduced an algorithm to calculate the semantic similarity between two short texts based on semantic nets and corpus statistics [57]. This algorithm uses information content provided by corpus statistics, such as Brown Corpus, to weigh and determine the importance of each word in each sentence. The integration of semantic similarity and order similarity are the basis of calculating the sentence similarity. The calculation of semantic similarity is based on two semantic vectors and information content, while the calculation of order similarity is based on two-word order vectors. Generating two semantic vectors and two words order vectors depend on a lexical database such as WordNet. The algorithm achieved acceptable Pearson correlation coefficient but it requires other performance measurements to compare it with other methods. Similarly, with our proposed model, we used a lexical database. However, Li et al. used a general lexical database such as WordNet, which may give irrelevant synonyms of the word's context, while we provide a lexicon and an ontology in a specific domain of interest in our model for semantic purposes. In addition, a drawback of Li et al.'s approach is that the generation of semantic vectors and word order vectors for two compared texts is only effective when the text length is short. As a result of that, this approach is not capable to generate the aforementioned vectors and compute the semantic similarity between two long texts.

Liu and Wang proposed a method for assessing text similarity based on an ontology [58]. This method semantically compares between two texts either a sentence or an entire document. The connections that exist between ontology concepts and text content are used to create concept vectors for either sentences or documents. In other words, sentences and documents' concept vectors are generated based on the concepts' weights in a sentence or a document. However, document concept vectors employ TF_IDF measure to determine words' importance in a specific document. Also, Liu and Wang used semantic similarity indexes, which are derived from the aforementioned connections, for searching and comparison purposes. A comparison is conducted of Liu and Wang method with other methods, such as word overlap, TF-IDF, and linguistic measures. As well, the authors computed four performance measurements in their evaluation, such as accuracy, precision, recall, and F-measure, and their method achieved higher values for all four measures compared to other methods.

Chahal et al. developed a semantic similarity model using relation measurements for web documents based on an ontology [59]. The model comprises of three main components, which are ontology processor, document processor, and calculation of semantic score components. The ontology processor component consists of concept and relation analyzer, which in turn is used to extract all concepts in a specific web document and relations between them. The document processor consists of a syntactic and a semantic analyzer, which in turn extract words using a specific dictionary in a specific domain. Then, the method calculates the semantic similarity between each pair of concepts in a document by finding the Relation space model (RSM) and the lexical patterns. The RSM has all relations of concepts in a document along with relations' frequencies. The comparison between two documents can be done by comparing the RSM of documents and calculating the semantic similarity score between two documents. The method

was evaluated on 50 documents and compared to the vector space model and Euclidean approaches and achieved better performance results compared with other approaches. However, this model needs to be run on a much larger dataset to measure semantic similarities between documents and obtain more performance results.

2.3 Summary

In this chapter, the first section provided a wide background on data loss and leakage prevention along with a comparison between several methods clarifying their goals, strengths, and weaknesses. The second section presented previous works related to insider threat detection and prediction, data loss and leakage prevention, and ontological search. The majority of insider threat detection and prediction work have depended on user activity and log analysis at the host or network level. In addition, some previous works on data loss and leakage prevention depend on TF-IDF metric to generate sensitive words' list. Some proposal used steganography to detect leaked data while some work applied machine learning techniques for detection.

In the next chapter, we introduce in detail our proposal for effective data loss prevention based on document semantic.

Chapter 3

Proposed Data Loss Prevention Model

In this chapter, we present our approach for generating and matching the DSS for data loss prevention. Figure 3.1 shows the block diagram covering the main steps of our proposed approach.

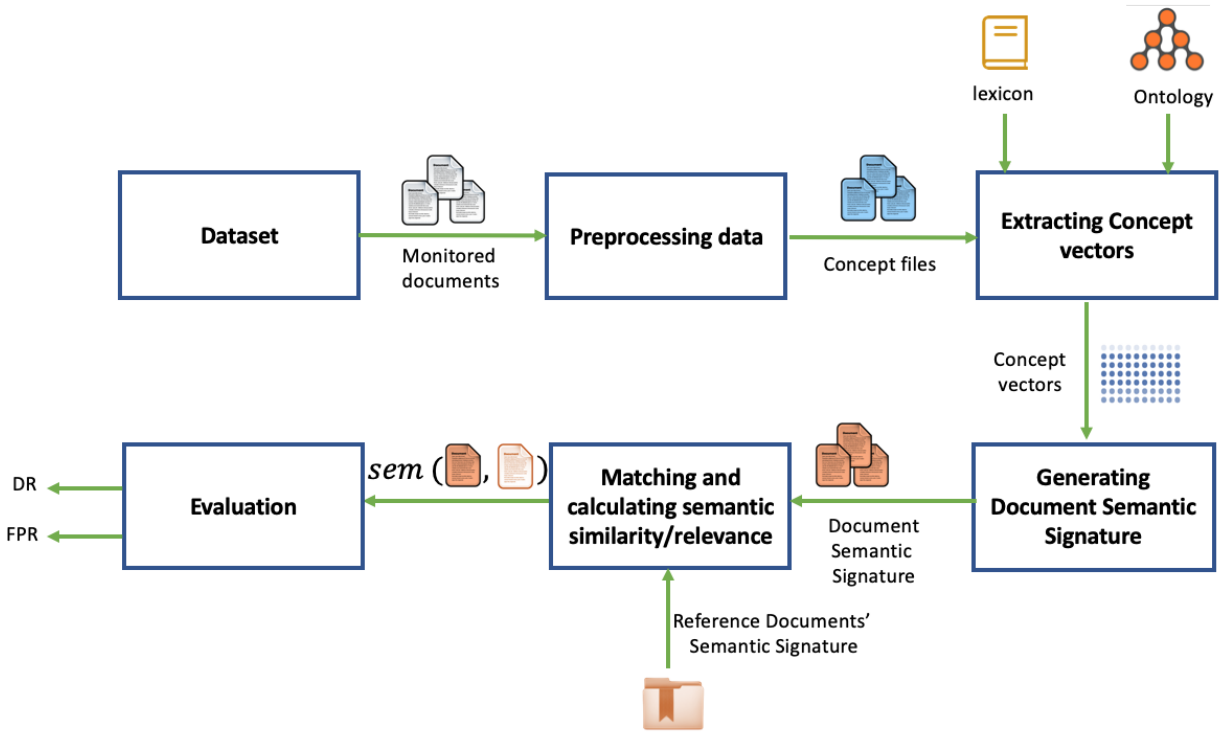


Figure 3.1 A block diagram of our proposed DLP model.

We will revisit the approach in more detail by starting with introducing the underlying conceptual components and then presenting how the DSS can be synthesized from these elements.

3.1 Ontology Concept Tree

An ontology is a formal representation in a hierarchical structure of a set of concepts and the relations between these concepts in a specific knowledge domain [38]. The ontology allows representing the concepts contained in a document. The relationships between these concepts provide the meaning of the content, also known as the semantic. Similar concepts or classes in the ontology are structured in a taxonomy structure referred to as *concept tree*. A *concept tree (CT)* describes the abstraction relationship (i.e. generalization/specialization) between similar concepts using a hierarchical structure. The root of the tree corresponds to the most abstract form of the concept, while intermediary nodes correspond to refined concepts, and leaf nodes correspond to instances.

An ontology may consist of several concept trees, each describing a group of related concepts. The collection of concept trees can be grouped in a larger tree representing the ontology. In this case an abstract node serves as the root of the large ontology concept tree. Commonly, such root node is referred to as *Thing*. It is more a placeholder or abstraction that allows bundling the different individual concept nodes in a single larger tree structure representative of the entire ontology.

Let $O = (C, R)$ denote the ontological tree for ontology O, where C and R correspond to the set of all concepts and the set of relationships among the concepts, respectively. By default C contains at least the abstract *Thing* as the root of the tree: $Thing \epsilon C$.

Given a concept node $c \in C$, let $ancestors(c)$ and $descendants(c)$ denote the set of all ancestor nodes and the set of all descendant nodes of c in the ontology tree, respectively.

We define the relevancy nodes for c as:

$$relevancy(c) = ancestors(c) \cup descendants(c) \cup \{c\} \quad (3\text{-}1)$$

The ***relevancy nodes*** for a specific concept node are the root node, the ancestor nodes, the descendant nodes and the node itself [29].

Let $depth(c)$ denote the depth of concept node $c \in C$ in the ontology, which represents the relative position of the node in the ontology concept tree, with respect to the root, as an integer value. The node's depth can be defined as the number of all nodes in the path from a specific concept node to the root (including itself and the root).

3.2 Concept File

The ontology, through the collection of concept trees, provides a generic characterization of the knowledge that needs to be protected. The specific knowledge (actual files, databases, etc.) that need to be protected are represented by their content semantics.

For a given document, we extract the most important concepts by applying several steps. The steps to extract the set of concepts to derive the concept file for a document are outlined as follows and depicted by Figure 3.2:

a) Preprocess document content by removing metadata. For instance, for emails, this involves removing email headers and keeping only email body.

b) Getting sentences: This function divides the file's text, which is extracted, e.g., from email's body, into separated sentences.

c) Removing stop words: This step filters out the sentences from the most common words in English based on stop word list [39].

d) Stemming: This function reduces the number of word types by deriving the roots.

e) Creating concept file: After applying the aforementioned sub-steps, this function saves the derived concepts in a text file.

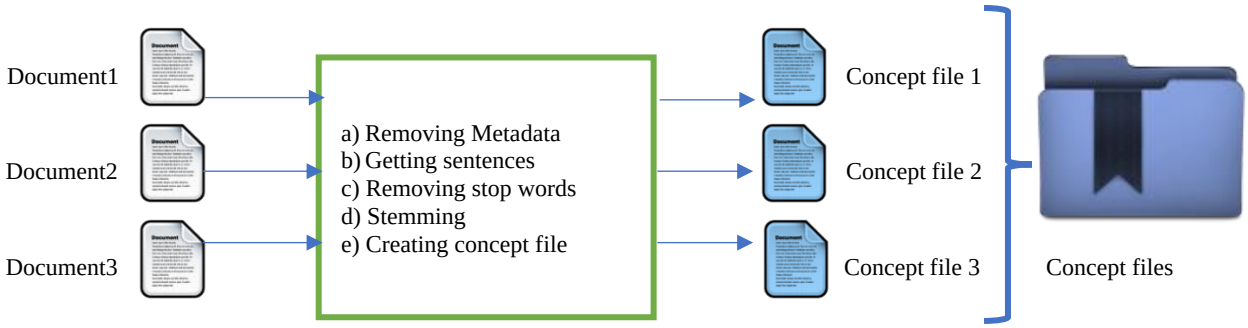


Figure 3.2 Steps for Extracting Concepts

3.3 Document Concept Tree

The document concept tree (DCT) captures the semantic of document content relative to a specific domain of knowledge represented by an ontology. Given a document, the DCT is constructed by extracting all the concepts from the document concept file that are available in the ontology. As part of this process, synonymous concepts are replaced by matching concepts available in the ontology [29].

Given an ontology $O = (C, R)$, the concept tree for a document d is defined as a triple $CT(d) = \{\{Thing\}, C_d, R_d\}$,

Where:

- $C_d = \{c_1, c_2, \dots, c_n\}$ is a set of concepts; each concept $c_i \in C_d$ is a word or phrase, and it is unique in C_d; also $C_d \subseteq C$.

- $R_d = \{r_1, r_2, \ldots, r_t\}$ is a set of relationships among concepts; each relationship $r_i \in R_d = (c_p, c_q, l_j), p \neq q, 1 \leq p, q \leq n, 1 \leq j \leq t$, connects two concepts $c_p, c_q \in C$. Label l_j is a term which labels relationship r_j.

Each document concept tree contains by default, as its root, "Thing", the root of the ontology. Algorithm 1 summarizes the steps for constructing the concept file for a given document, as shown below.

Algorithm 1 Summarizing Document and Extracting Concept Map File

```
/* arrList is an ArrayList of String */
/* file is a text File */
/* ConceptMap is a text file which has a set of extracted concepts from file*/
Input: void
Output: void
1: procedure SUMMARY( )
2:     File file ← loadFile(filepath);
3:     arrList ← setdocument(file, arrList);
4:     arrList ← GetSentences(file);
5:     arrList ← removestopwords(arrList);
6:     Tokenization();
7:     Stemming();
8:     Significant();
9:     File ConceptMap ← GetKeywordsToOutputFile();
10: end procedure=0
```

The runtime complexity of Algorithm 1 is $O(n\alpha^2)$, where *n* is the total number of concepts in a document, and α is the total number of sentences in a file. Its space complexity is $O(\alpha + n^2)$. The time and space complexity of Algorithm 1 are both quadratic functions. The runtime and memory requirements for Algorithm 1 will be increased gradually based on the increasing number of sentences in a file and number of concepts in a document, respectively.

3.4 Document Semantic Signature

Given a document d, the semantic signature of the document $SS(d)$ captures objectively the relevancy of each of the nodes c_i of the document concept tree with respect to each of the nodes c_j of the ontology concept tree. It is defined as the following matrix:

$$SS(d) = [v_{ij}]_{\substack{1 \le i \le n \\ 1 \le j \le m}} \qquad (3\text{-}2)$$

$$Where \quad v_{ij} = \begin{cases} depth(c_i) & if\ c_j \in relevancy(c_i) \\ 0 & otherwise \end{cases} \qquad (3\text{-}3)$$

and *n* and *m* correspond to the total number of concepts in the document and ontology concept trees, respectively.

The row in the *SS(d)* matrix are referred to as concept vectors, i.e., row i $(1 \le i \le n)$ corresponds to the concept vector for concept c_i.

Figure 3.3 depicts the generation of the semantic signature for different documents.

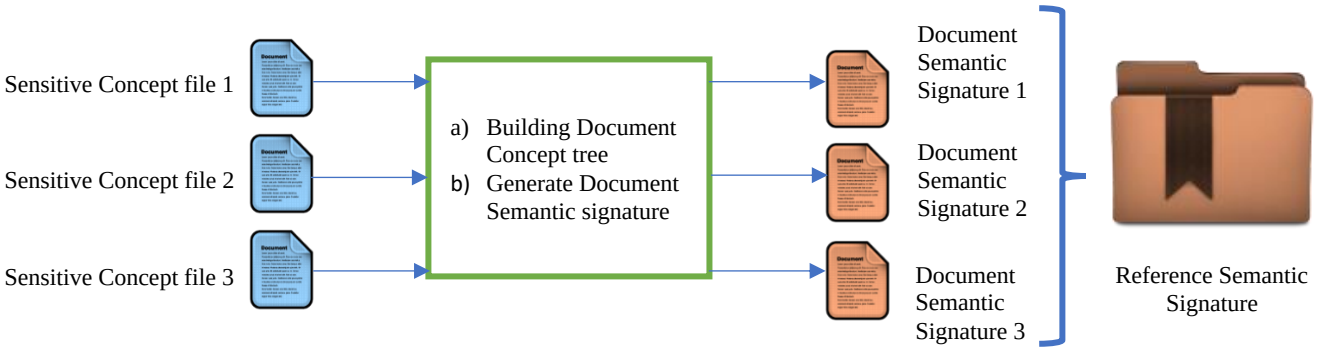


Figure 3.3 Generating Reference Semantic Signature

3.5 Semantic Signature Matching

Given a set of document $M = (d_1, \dots, d_x)$ considered sensitive that are being protected, we extract from each of the documents their semantic signature. Let n_{d_i} denote the number of concepts involved in the concept file of document d_i. So, each $SS(d_i)$ is a $n_{d_i} \times m$ matrix.

The set of semantic signatures represent the *reference signature* $SS(M) = \big(SS(d_1), \dots, SS(d_x)\big)$.

Note that the matrices corresponding to the semantic do not have necessarily the same number of rows, as the number of concepts may be different in each document. In contrast, they all have the same number of columns *m*.

Given a suspected document *d*, data loss prevention consists of checking for similarity against the protected documents. This takes place by comparing the semantic signature $SS(d)$ against the reference signature $SS(M)$.

The matching consists of tracking the occurrence of each of the concept vectors of the monitored or suspicious document *d* in each of the semantic signatures in the reference signature. Given *i* ($1 \leq i \leq$ n), concept vector $v_i = [v_{ij}]_{\substack{1\leq i\leq n \\ 1\leq j\leq m}}$ from $SS(d)$ occurs in semantic signature $SS(d_k)$ from $SS(M)$, if one of the rows in $SS(d_k)$ matches exactly v_i. The matching involves initially calculating and aggregating the cosine similarity (CS) between the concept vectors of the documents as

$$CS\big(SS(d)\,, SS(d_k)\big) = \frac{\sum_{l=1}^{n}\sum_{r=1}^{n_{d_k}} CS(v_l(d), v_r(d_k))}{n \times n_{d_k}} \qquad (3\text{-}4)$$

where $CS(v_l(d), v_r(d_k))$ is the cosine similarity between the l_{th} concept vector of $SS(d)$ and the r_{th} concept vector of $SS(d_k)$, respectively, which is defined as follows:

$$CS(v_l(d), v_r(d_k)) = \frac{\sum_{j=1}^{m} v_{lj}(d) \times v_{rj}(d_k)}{\sqrt{\sum_{j=1}^{m} v_{lj}(d)^2} \times \sqrt{\sum_{j=1}^{m} v_{lj}(d_k)^2}} \qquad (3\text{-}5)$$

Then, let δ_{ik} denote the matching outcome, defined as follows:

$$\delta_{ik} = \begin{cases} 1 & if\ CS\big(SS(d), SS(d_k)\big) = 1 \\ 0 & otherwise \end{cases} \qquad (3\text{-}6)$$

Based on an ontology concept tree, we calculate the semantic measurement between two documents using the semantic similarity and/or semantic relevance metrics.

The calculation of the semantic similarity between two concepts depends on the concepts' content similarity while the calculation of the semantic relevancy depends on the existence of the relations between those two concepts. To illustrate, semantic similarity measures the content similarity of two concepts in an ontology when they have a taxonomic relation between them such as is-A, subclass, or is-a-child relation. On the other hand, semantic relevancy measures the relatedness between two concepts in an ontology when they have explicit or implicit relations and it is not necessary to have similar contents.

We define in the subsequent subsections the semantic similarity and semantic relevance metrics used in our work.

3.6 Semantic Similarity metrics

To calculate the similarity of the monitored document with respect to the reference document, we use one of the two ontology-based semantic similarity metrics defined in the following. Later, in the experimental evaluation chapter, we will compare these two metrics and assess their adequacy for our model.

3.6.1Ontology-based Semantic Similarity Metric 1

In this sub-section, we calculate the similarity between the monitored and the reference documents by applying a combination of two similarity metrics: a simple frequency model and the Jaccard index.

In the frequency model, individual matching frequencies are determined and stored in a vector $F = [f_k]_{1 \leq k \leq x}$

Where:

$$f_k = \frac{\sum_{i=1}^{n} \delta_{ik}}{n_{d_k}} \qquad (3\text{-}7)$$

In the Jaccard model, indices are calculated by comparing the monitored document signature against each of the reference document signatures, using the approach outlined above, whereby the number of matching concept vectors is tracked. The outcome of the comparisons is provided in a vector $J = [J_k]_{1 \leq k \leq x}$

Where:

$$J_k = \frac{\sum_{i=1}^{n} \delta_{ik}}{n_{d_k} + n - \sum_{i=1}^{n} \delta_{ik}} \qquad (3\text{-}8)$$

Each of the similarity metrics are compared against separate predefined thresholds to establish similarity or dissimilarity.

Let th_f and th_J denote the thresholds for the frequency and Jaccard metrics, respectively. The monitored document d is suspected to contain portion of some of the reference documents if:

$$\exists i \epsilon \{1, \dots, x\} \; such \; that \; \left(\left(J_k \geq th_J \right) and \left(f_k \geq th_f \right) \right)$$

Algorithm 2 depicts the steps for extracting the concept tree of a document and measuring the similarity between a specific document and sensitive ones. To illustrate, algorithm 2 consists of loading several files including reference documents' semantic signature, monitored textual documents, ontology files, and a synonym file in a specific domain of interest. Also, it includes creating an ontology concept tree and a document concept tree. As well, it compares concept vectors of monitored and reference documents and calculates the semantic similarity metrics including the simple frequency and Jaccard metrics.

The runtime complexity of Algorithm 2 is $O(m^3x + q)$, whereas its space complexity is $O\big((n + x)m\big)$, where q is the total number of ontology files, m is the total number of concepts in the ontology, x is the total number of sensitive documents, and n is the total number of concepts in the document. The time complexity is cubic, while the space complexity is linear. The total running time for Algorithm 2 is increased basically with increasing the number of concepts in the ontology. Also, the total memory requirements are increased linearly with the increase of the input size including the number of sensitive documents and the number of concepts in the document and ontology.

Algorithm 2 Extracting Document Concept Tree and Measuring Similarity

Input: void

Output: void

```
1: procedure ExtractTreeMeasureSimilarity()
2:     SynonymLoader();
3:     LoadRDFOntology();
4:     File TestFile ← LoadTestFile(TestFilePath);
5:     String FName⟵ TestFile.getName();
6:     Ontology.search(FName);
7:     extractDocConceptTree(FName);
8:     printVectors(OntConceptVectorPath, OntDocVectorPath);
9:     calculateOntDocVecSize();
10:    printVectors(ExtConceptVectorPath, ExtDocVectorPath);
11:    calculateExcDocVecSize();
12:    List < File > RefFiles ⟵ LoadRefDataset();
13:    int DocIndex ⟵ 1;
14:    for each File RefFile : RefFiles do
15:        getDocVectorFromFile(RefFile.getCanonicalPath(), DocIndex);
16:        comparingConcepts(DocIndex);
17:        MeasureSimThreshold(DocIndex);
18:        MeasureConceptsSim(DocIndex);
19:        MeasureJaccardSim(DocIndex);
20:        DocIndex + +;
21:    end for
22:    printSimResults();
23: end procedure=0
```

Figure 3.4 shows a sample text document, specifically an email from the Enron email dataset that will be described in details in the dataset chapter [40] [41]; let's refer to it as reference (or sensitive) document d_1.

```
Message-ID: <7879273.1075840857246.JavaMail.evans@thyme>
Date: Tue, 22 May 2001 19:37:00 -0700 (PDT)
From: edhearst@earthlink.net
To: louise.kitchen@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Edward Hearst  <edhearst@earthlink.net>
X-To: louise.kitchen <louise.kitchen@enron.com>
X-cc:
X-bcc:
X-Folder: \ExMerge - Kitchen, Louise\'Americas\HR
X-Origin: KITCHEN-L
X-FileName: louise kitchen 2-7-02.pst

Dear Ms. Kitchen,

I am currently working as a V.P. at Commerce One managing the Global Trading
Web.  I have followed Enron's e-commerce activities for some time.  I am
currently exploring other career opportunities and thought I might be able
to contribute to Enron's online initiative.  I previously worked at the
House Energy and Commerce Committee, the FCC, the State Dept, and Jones, Day
Reavis & Pogue.  I not only have experience in the B2B world through my work
at Commerce One, but also in e-commerce generally, trade and
telecommunications through these other positions.  This combination of
experience could be helpful to Enron in expanding your B2B effort, including
governance and interoperability issues, developing new business, and in
dealing with domestic and international regulatory matters related to
e-commerce.

A copy of my resume is attached.  If you have a few moments to talk, I would
greatly appreciate it.

Thanks and best regards,
```

Figure 3.4 A sample document from Enron email dataset

The extracted concept file, document concept tree, and document semantic signature from the above email sample are shown in Figure 3.5.

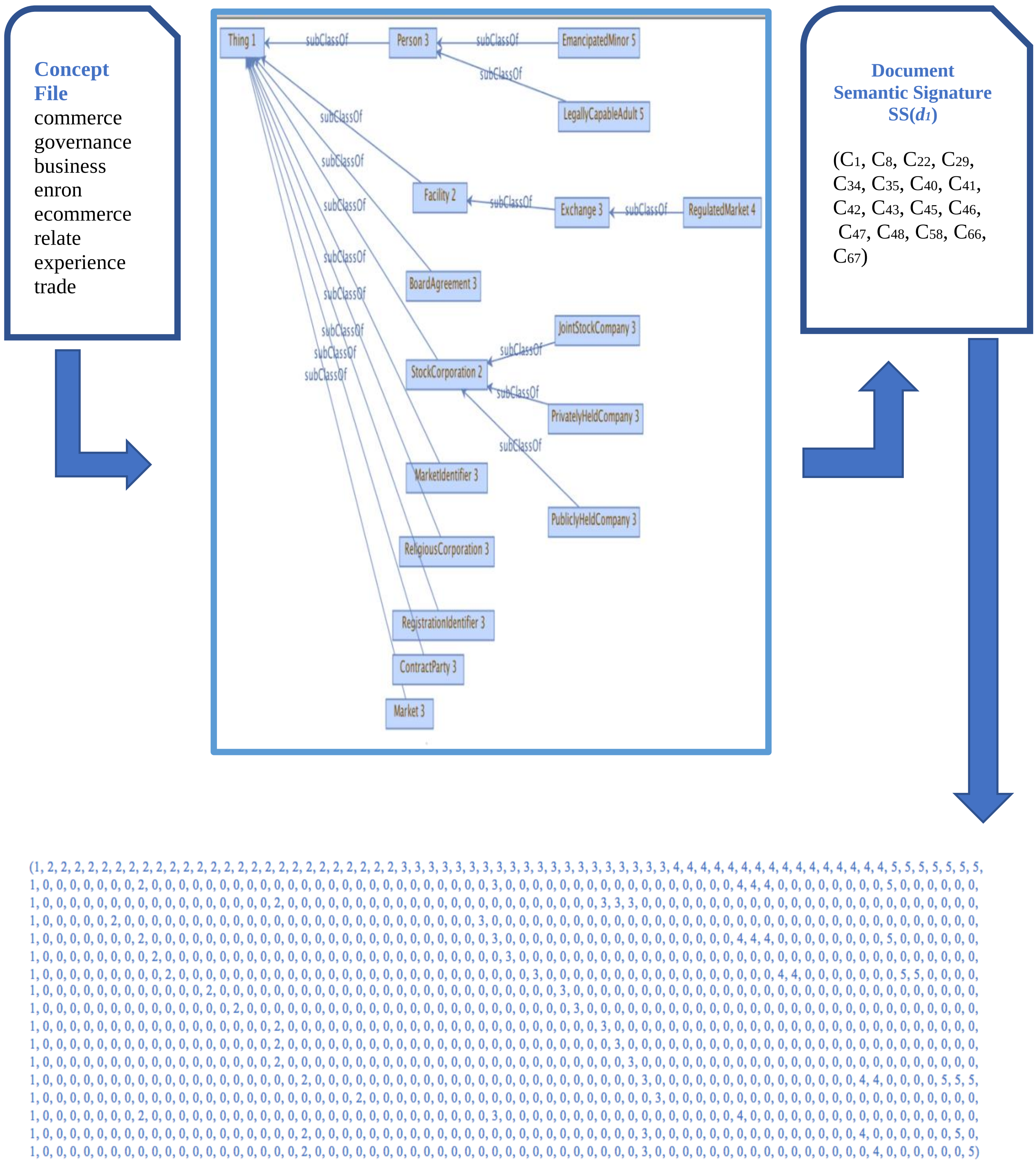


Figure 3.5 Generating Concept tree and Document Semantic Signature SS(d1) for Reference Document d1

In order to illustrate the matching process, let's assume that we have three sensitive documents in the reference (including the email sample given above, d_1): $M = (d_1, d_2,\ d_3.)$

Using the above approach, we can generate the reference signature $SS(M) = \{SS(d_1), SS(d_2), SS(d_3)\}$. Figure 3.6 and Figure 3.7 illustrate the generation of the semantic signatures for the 2 additional documents.

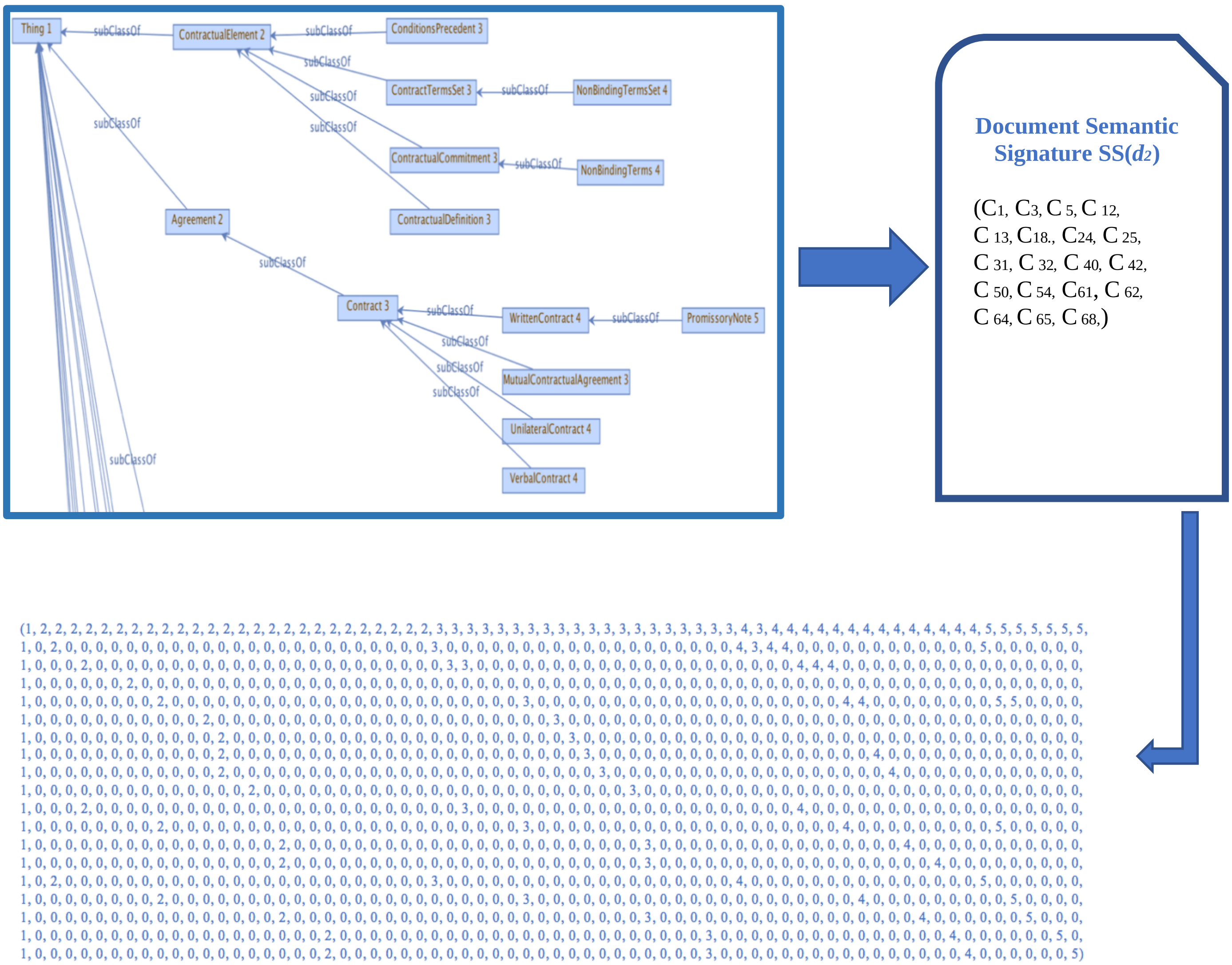
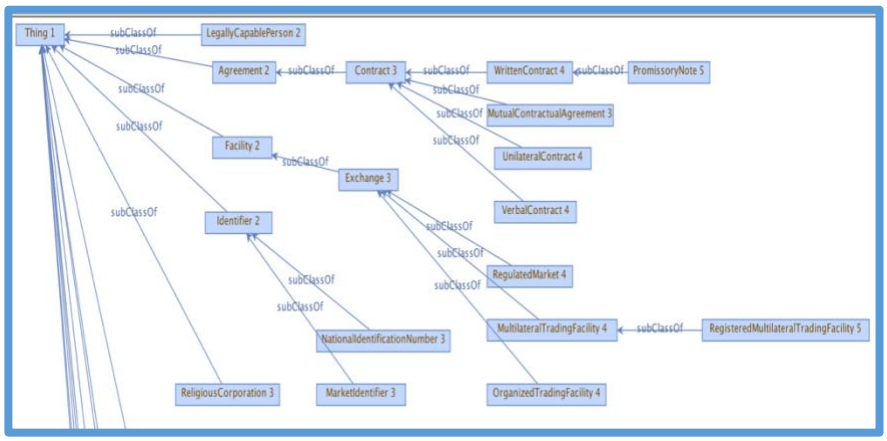


Figure 3.6 Generating concept tree and Document Semantic Signature SS(d2) for Reference Document d2



Document Semantic Signature SS(d_3)

(C_1, C_6, C_9, C_{10}, C_{13}, C_{14}, C_{19}, C_{22}, C_{23}, C_{28}, C_{29}, C_{30}, C_{32}, C_{33}, C_{34}, C_{37}, C_{38}, C_{41}, C_{42}, C_{43}, C_{44}, C_{45}, C_{48}, C_{49}, C_{50}, C_{51}, C_{52}, C_{57}, C_{58}, C_{59}, C_{64}, C_{65}, C_{66}, C_{68}, C_{70})

(1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5,
1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2,)

Figure 3.7 Generating concept tree and Document Semantic Signature SS(d3) for Reference Document d3

Let's assume that we have two monitored documents CF_1 and CF_2, that need to be matched against the reference signature. Figure 3.8 shows the matching process of monitored document CF1 against the reference signature $SS(M)$.

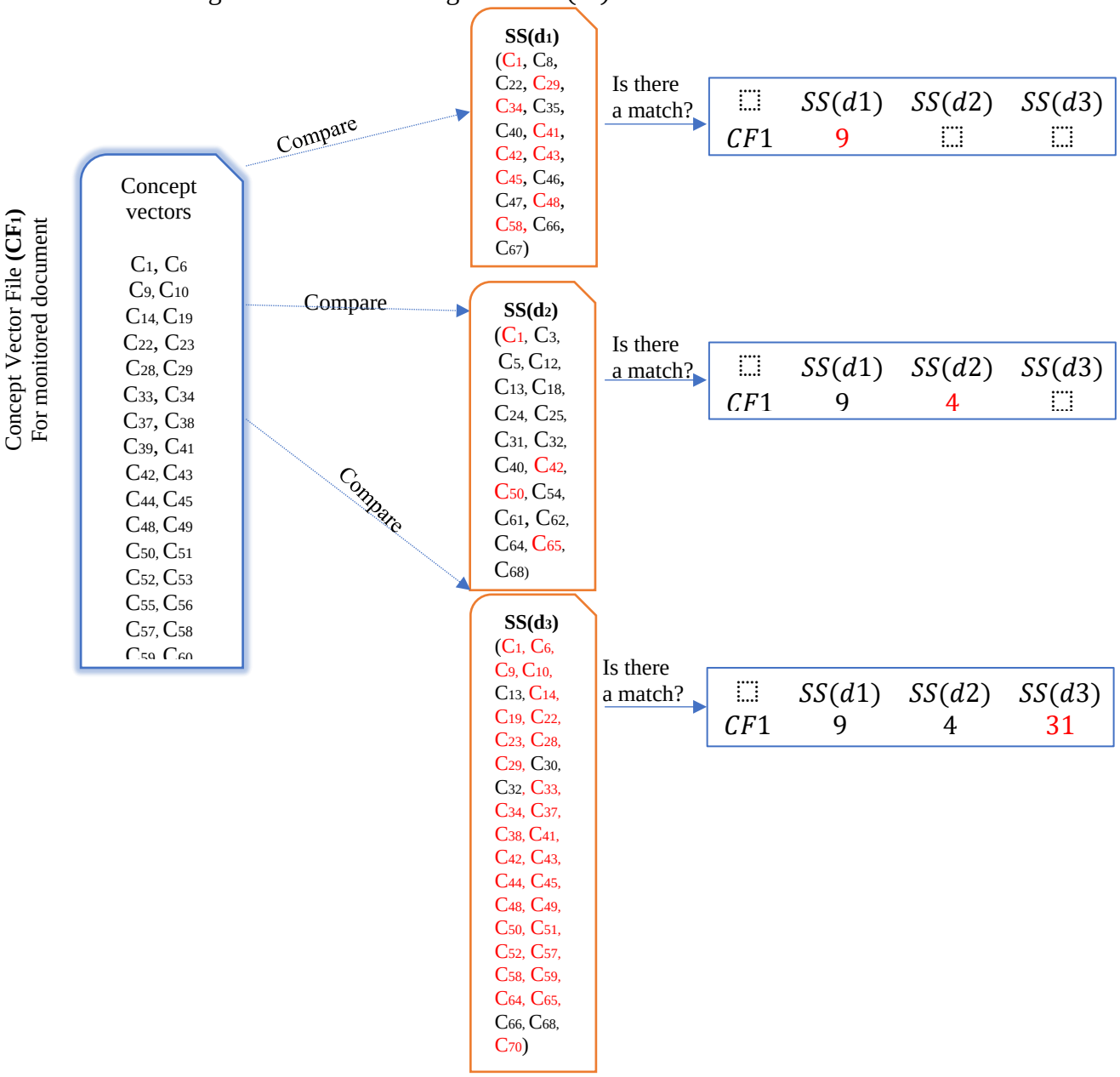


Figure 3.8 Matching process of monitored document CF1 against reference signature.

This figure shows the comparison process of each concept vector in the monitored document CF1 against all sensitive documents' semantic signatures. If there is a match between monitored concept vector and document semantic signature, then the frequency will be incremented by one and saved to the frequency matrix and so on. Then, the same steps will be repeated for all concept vectors in the monitored document against the remaining documents semantic signatures. As an example, CF1 has 9 matched concepts in sensitive document SS(d1), 4 matched concepts in SS(d2), and 31 matched concepts in SS(d3).

Figure 3.9 below illustrates the comparison of monitored document CF2 against the reference signature.

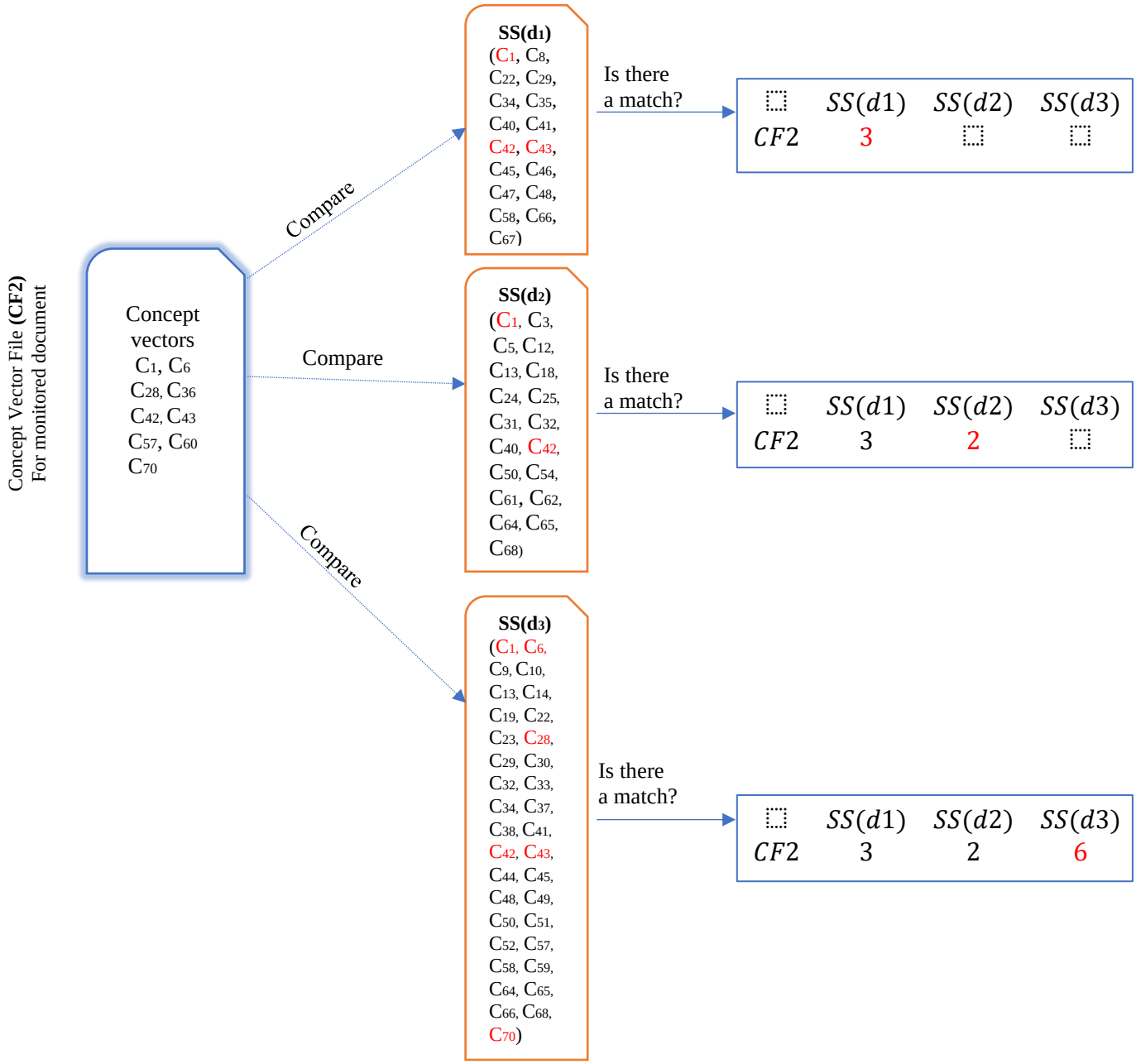


Figure 3.9 Matching process of monitored document CF2 against reference signature.

This figure shows the comparison process of each concept vector in the monitored document CF2 against all sensitive documents semantic signatures. If there is a match between monitored concept vector and document semantic signature, then the frequency will be incremented by one and saved to frequency matrix and so on. Then, the same steps will be repeated for all concept vectors in the monitored document against the remaining documents semantic signatures. As an example, CF2 has 3 matched concepts in sensitive document SS(d1), 2 matched concepts in SS(d2), and 6 matched concepts in SS(d3).

Figure 3.10 below shows the general steps for calculating the frequency for each monitored concept vector file.

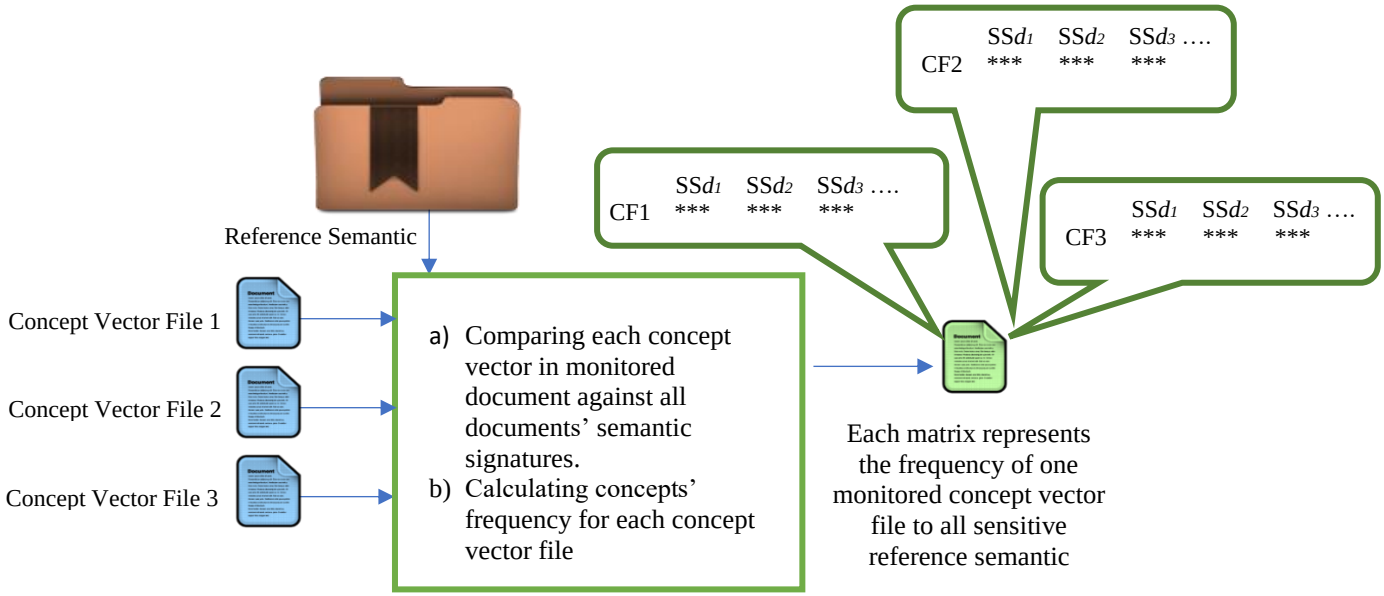


Figure 3.10 Comparing vectors and calculating concepts' frequency

The two matrices below represent the matching frequencies for the two monitored documents CF1 and CF2.

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$CF1$	9	4	31

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$CF2$	3	2	6

Also, the two matrices below show the frequency percentage for monitored concept vector files CF1 and CF2, which show that the highest percentage of frequency of CF1 is 91.18% in SS(d_1), while the lowest frequency percentage is 22.22% in SS(d_2). On the second monitored file CF2, the highest frequency percentage is 20% in SS(d_1), while the lowest is 11.11% in SS(d_2).

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$F(CF1)$	60%	22.22%	91.18%

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$F(CF2)$	20%	11.11%	17.65%

In addition, the Jaccard index is calculated below for both monitored documents. The two matrices below show that the highest Jaccard index for CF1 is 75.96% while the highest for CF2 is 42.86%.

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$J(CF1)$	21.43%	8%	79.49%

	$SS(d1)$	$SS(d2)$	$SS(d3)$
$J(CF2)$	42.86%	8%	16.22%

From the measures above, our model will classify CF1 as a suspicious file since the Frequency F(CF1) = 91.18% against SS(d3) which is higher than 75% and the Jaccard index J(CF1) = 75.96% against SS(d3), which is higher than 75%, too.

3.6.2 Ontology-based Semantic Similarity Metric 2

In our model, we applied another similarity metric to measure the semantic similarities between two concepts based on an ontology. The metric was originally defined in the work of Saad et al. [42]. The semantic similarity between two concepts in an ontology is measured based on the commonalities and the differences between them. In other words, the relations between two concepts and their lowest common ancestor in an ontology capture the commonalities between these two concepts. In contrast, the locations of two concepts in an ontology capture the differences between them. Given two concepts c_1 and c_2 belonging to documents $d_1$and d_2, the semantic similarity between these concepts is defined as follows:

$$sim\ (c_1, c_2) = 1 - \frac{(path(c_1, LCA\ (c_1, c_2)) + path(c_2, LCA\ (c_1, c_2)))}{(depth(c_1) + depth(c_2))} \qquad (3\text{-}9)$$

Where $path(c_1, LCA\ (c_1, c_2))$ is the length of the shortest path from concept c_1 to the least common ancestor LCA of c_1 and c_2, and $depth(c_1)$ is the depth of concept c_1 in the

concept tree. The output of $sim\ (c_1, c_2)$ is a value between [0,1] where 1 corresponds to exact match while 0 corresponds to no match between the two concepts.

We define the document semantic similarity by summing and averaging the semantic similarities between their respective concepts, as defined as follows:

$$sim\ (d_1, d_2) = \frac{\sum_{c_1 \epsilon\, d_1, c_2 \epsilon\, d_2} sim(c_1, c_2)}{n_{d_1} \times n_{d_2}} \qquad (3\text{-}10)$$

Where n_{d_1} and n_{d_2} are the total number of concepts in d_1 and d_2, respectively.

3.7 Ontology-based Semantic Relevance Metric

To capture the relevancy between concepts, in our model, we use a semantic relevance metric that depends on the relations between concepts in an ontology.

3.7.1 Ontology Relations

Given an ontology $O = (C, R)$, the concept tree for a document d is defined as triple $CT(d) = \{\{Thing\}, C_d, R_d\}$. There are two types of relations r ∈ R between any two concepts in an ontology:

1- Explicit relations $r_{explicit}$, which are predefined between classes and need to be extracted from the ontology.

2- Implicit relations $r_{implicit}$,which are hidden and need to be inferred by some inference rules.

In our model, we extract the relations between the classes whether they are explicit or implicit relations. Any two concepts in the ontology are linked in one of three different ways: they have no relation, one relation, or more than one relation. In our implementation, we used the *Jena* library for ontology reasoning to infer implicit relations.

Given two documents $d_1$and d_2, the set of relations between these documents are the combination of all extracted explicit and implicit relations between their respective classes, defined as followed:

$$R(d_1, d_2) = \bigcup_{x \in C_{d_1}, y \in C_{d_2}} [r_{explicit}(x, y) \cup r_{implicit}(x, y)] \qquad (3\text{-}11)$$

where $r_{explicit}(x, y)$ and $r_{implicit}(x, y)$ denote the sets of explicit and implicit relations between each two concepts *x* and *y*. Also, let R_{xy} denote the set of all relations between *x* and *y*, including explicit and implicit ones:

$$R_{xy} = r_{explicit}(x, y) \cup r_{implicit}(x, y) \qquad 3\text{-}12)$$

One of the significant steps to calculate the semantic relevance between two documents is to extract all explicit and implicit relationships between concepts based on an ontology. To illustrate, there are some concepts in the FIBO ontology (introduced later) that have just a "is subclass of" relation as shown in Table 3.1, while other concepts have more than one relation as extracted and shown in Table 3.2.

Table 3.1 Examples of extracted explicit relations between different concepts in FIBO ontology.

Ontology module	**Class name**	**Relation**	**Class name**	**Ontology module**
People.rdf	PostCodeArea	is subClass of	PhysicalLocation	People.rdf
Markets.rdf	Commerce	is subClass of	CommercialActivity	Markets.rdf
Contracts.rdf	MutualCommitment	is subClass of	Commitment	Contracts.rdf

Table 3.2 shows examples of extracted explicit and implicit relations between two concepts, in which the ontology module and class name of the two concepts along with the extracted relations are specified.

Table 3.2 Examples of extracted explicit and implicit relations between two concepts in FIBO ontology. The 1st row is an explicit relation while the remaining rows are implicit relations.

Ontology module	Class name	Relation	Class name	Ontology module
People.rdf	BirthCertificate	is subClass of	IdentityDocument	People.rdf
People.rdf	BirthCertificate	hasExpirationDate	IdentityDocument	People.rdf
People.rdf	BirthCertificate	hasDateOfIssuance	IdentityDocument	People.rdf
People.rdf	BirthCertificate	hasUniqueIdentifier	IdentityDocument	People.rdf
People.rdf	BirthCertificate	verifiesPlaceOfBirth	IdentityDocument	People.rdf
People.rdf	BirthCertificate	verifiesAddress	IdentityDocument	People.rdf
People.rdf	BirthCertificate	Identifies	IdentityDocument	People.rdf
People.rdf	BirthCertificate	isIssuedBy	IdentityDocument	People.rdf
People.rdf	BirthCertificate	verifiesDateOfBirth	IdentityDocument	People.rdf
People.rdf	BirthCertificate	isAbout	IdentityDocument	People.rdf

As well in Sport ontology (introduced later), there are some concepts that have just a "is subclass of" relation as shown in Table 3.1, while other concepts have more than one relation as extracted and shown in Table 3.2.

Table 3.3 Examples of extracted explicit relations between different concepts in Sport ontology.

Ontology module	Class name	Relation	Class name	Ontology module
sport.owl	InterdependentTeam	is subClass of	Team	sport.owl
sport.owl	GroupSystem	is subClass of	RankingSystem	sport.owl
sport.owl	SwissSystem	is subClass of	GroupSystem	sport.owl
sport.owl	Sport	is subClass of	Activity	sport.owl

Table 3.4 Examples of extracted explicit and implicit relations between two concepts in Sport ontology. The 1st row is an explicit relation while the 2nd row is an implicit relation.

Ontology module	Class name	Relation	Class name	Ontology module
sport.owl	Match	is subClass of	SportsEvent	sport.owl
sport.owl	Match	hasPart	SportsEvent	sport.owl

3.7.2 Semantic Relevance metric

By capturing the semantic relevancy, we can measure the relatedness between a group of concepts or a group of instances in the ontology.

Given two concepts *x* and *y* belonging to documents $d_1$and d_2, the semantic relevancy between these concepts is defined as follows:

$$sem_{rel}(x,y) = \frac{\sum_{r \epsilon R_{xy}} w(r)}{\underset{u \epsilon C_{d_1}, v \epsilon C_{d_2}}{Arg\ max\ (R_{uv})}} \qquad (3\text{-}13)$$

where *w(r)* denotes the weight of a relation $r \in R_{xy}$ and $\underset{u \epsilon C_{d_1}, v \epsilon C_{d_2}}{Arg\ \max\ (R_{uv})}$ is the maximum number of relations between any two concepts in the ontology. We assume that each relation has a weight value equal to 1.

We define the document semantic relevancy by summing and averaging the semantic relevancies between their respective concepts, as defined as follows:

$$sem_{rel}(d_1,d_2) = \frac{\sum_{x \in C_{d_1}, y \in C_{d_2}} sem_{rel}\,(x,y)}{n_{d_1} \times n_{d_2}} \qquad (3\text{-}14)$$

Where n_{d_1} and n_{d_2} are the total number of concepts in d_1 and d_2, respectively.

3.8 Summary

In this chapter, we introduced our data loss prevention model by defining the underlying components, and by explaining, how these components help generate the DSS. We explained how using semantic similarity and relevance metrics, data leak can be monitored by comparing a sample document against reference sensitive documents.

In the next chapter, we will present the datasets and ontologies used in our experiments, and on this basis we will evaluate the effectiveness of the DSS model.

Chapter 4

Datasets, Ontologies, and Experiments

The backbone of the DSS model are the ontologies that capture domain-specific knowledge conveyed by the documents being monitored for data leak. To evaluate our model, we need datasets that contain real-world data leaks and the associated ontologies. In this chapter, we present two different datasets and corresponding ontologies that were used to validate experimentally the DSS model. Using the datasets, we conduct various experiments to evaluate the effectiveness of the proposed approach.

4.1 Datasets and Ontologies

To evaluate our model, we need a dataset that clearly identifies data that can be categorized as sensitive and potentially leaked information from this dataset. Unfortunately, there is no publicly available dataset that fully addresses this requirement. Hart et al. used for their experiments five datasets including DynCorp, TM, Mormon, Enron Emails, and Google private documents [30]. Despite their claim that the datasets would be available publicly, after reaching out, they couldn't share the datasets, claiming privacy restrictions.

The ontology is a data model to represent a set of concepts and the relations between them in a specific domain of interest in a hierarchical structure. The ontology is used mainly for information representation, organization, and reasoning in a particular domain. Also, the

concepts in an ontology may have explanation and definition to interpret concepts' usage in a particular ontology.

In our research, we used two different datasets and ontologies, which are related to business and sports domain of interest.

4.1.1 Business Dataset and Ontology

4.1.1.1 Enron Email Dataset

In our work, we used a real-life dataset, specifically a subset of the Enron email dataset focusing on business activities [41]. We grouped the emails by threads, each thread consisting of an initial email, and corresponding reply and forwarded messages. When a message is classified as sensitive, it is commonplace to categorize messages belonging to the same thread as sensitive. For instance, the initial message can be categorized as confidential by the sender, and then any other follow-up messages (part of the thread) would be treated as such. Under such a scenario, data loss prevention would consist of flagging the original messages as sensitive and monitoring follow-up messages for possible leakage.

We used a similar view in structuring the aforementioned email dataset to evaluate our proposed model. As mentioned above, we focused on only Enron Business emails which are 447 emails. We grouped these emails into threads based on the initial email, the responses, and the forwarded emails, yielding in total 375 threads. Figure 4.1 shows the total number of threads based on the number of emails. Each thread has either one, two, three, or four emails.

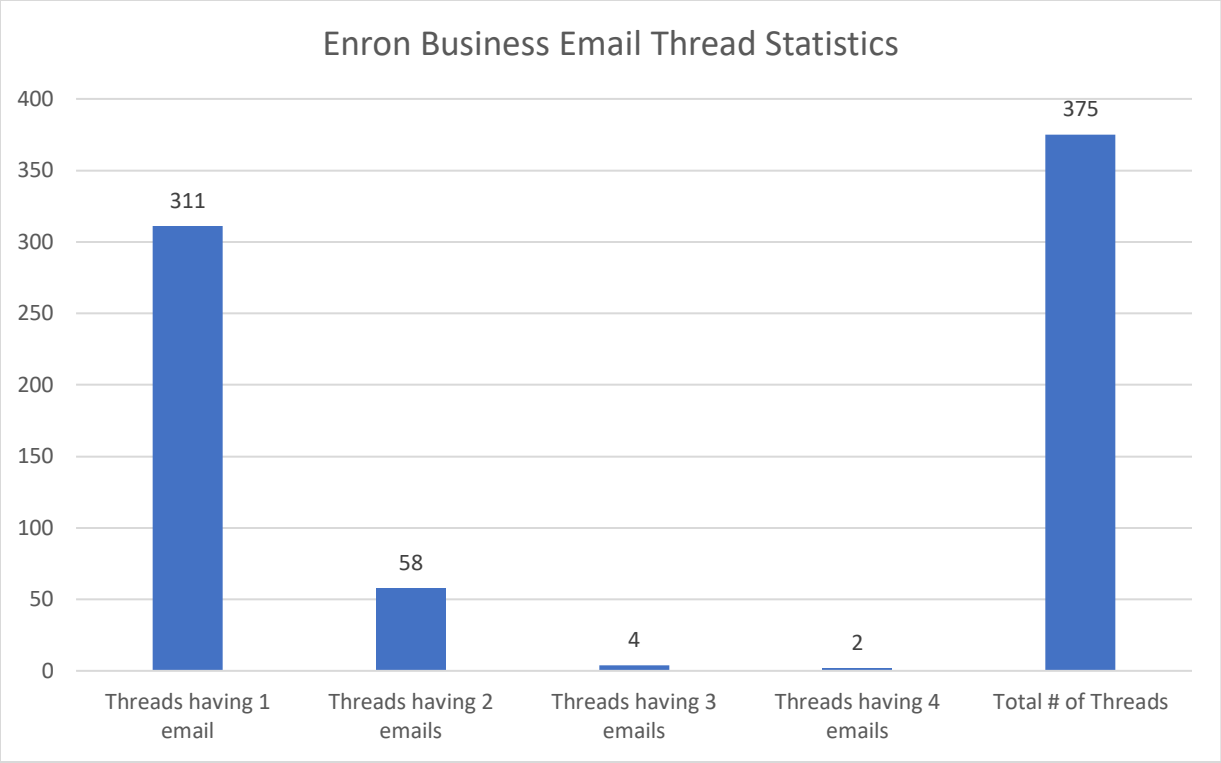


Figure 4.1 Breakdown of number of threads with different number of emails

A sample email of Enron email dataset is shown in Figure 4.2. The goal of the evaluation is to test the ability of the proposed model to detect leakage of information classified as sensitive while ignoring non-sensitive information. For a thread to be usable in testing leakage of sensitive information, we need at least 2 samples in it, so as to be able to use at least one sample as a reference and the remaining samples for testing. Under such constraint, we treated all the threads with a single message as non-sensitive, while the remaining threads were treated as sensitive.

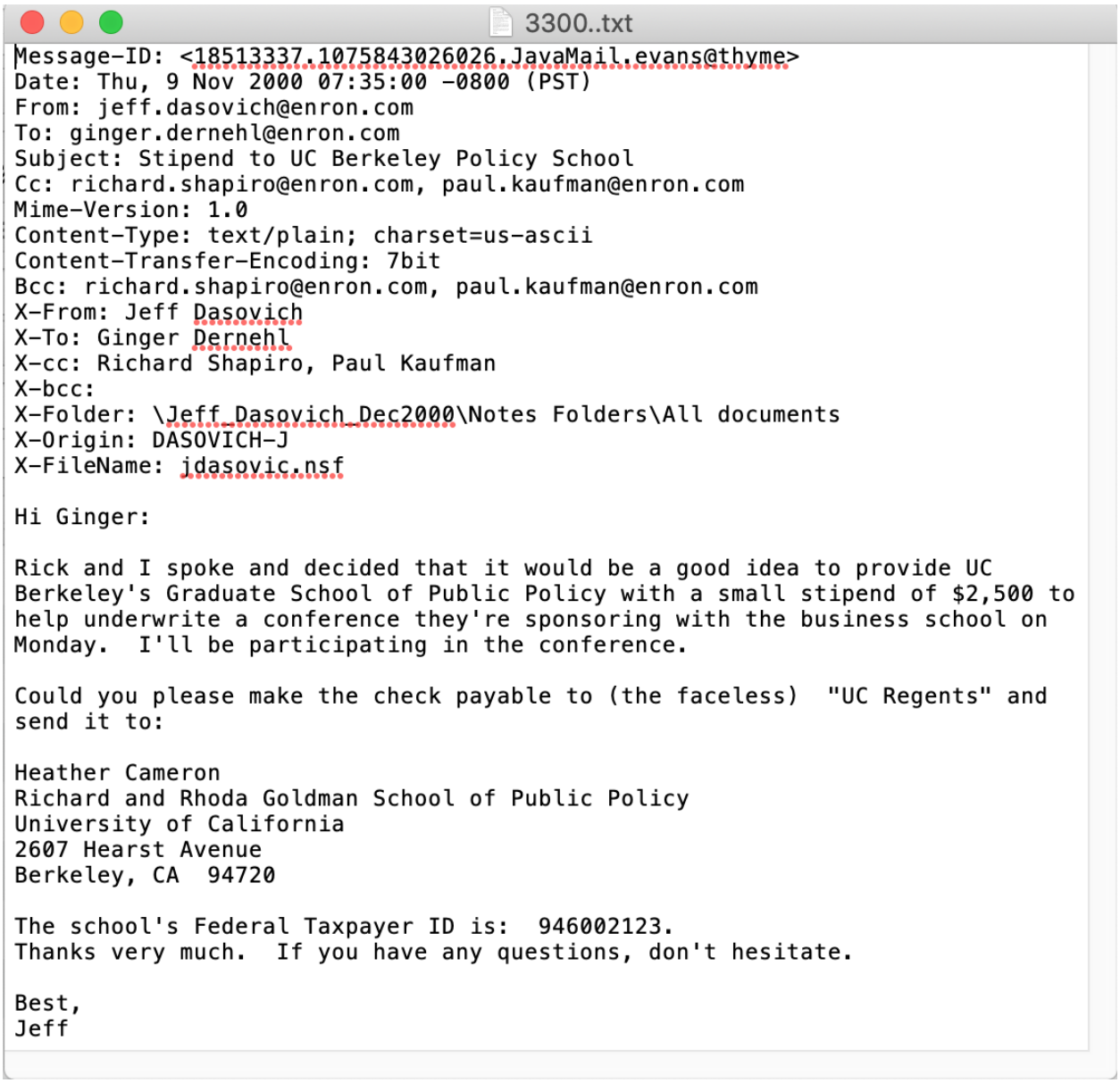


Figure 4.2 A sample email of Enron email dataset

4.1.1.2 Financial Industry Business Ontology

Our approach depends on using an ontology that describes the domain of knowledge being protected. In accordance with the domain covered by our selected dataset, we chose the Financial Industry Business Ontology (FIBO), which consists of 11 core domains, 49 modules and 418 ontology files (FIBO) [43]. Since FIBO ontology is a huge ontology, we have chosen People, Corporations, Markets, and Contracts modules as a partial representation of the FIBO ontology to implement our model. Figure 4.3 shows a partial representation of four modules of FIBO concept tree. Furthermore, business synonyms keywords' lexicon has been provided in our proposed model to assist in finding semantic keywords [44].

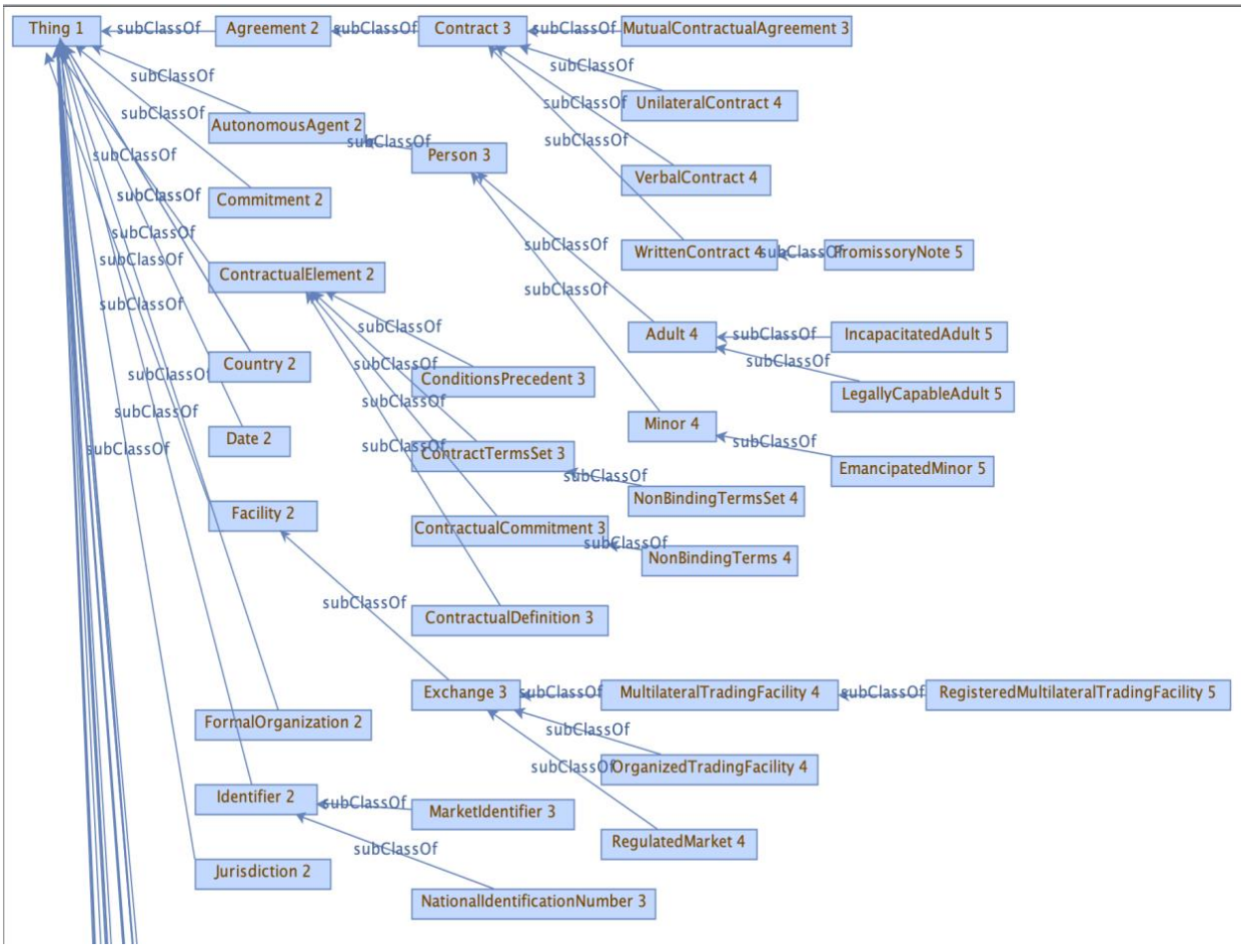


Figure 4.3 Partial representation of FIBO ontology

Table 4.1 shows the concepts from the partial FIBO ontology graph in Figure 4.3, along with their node labels and depths.

Table 4.1 Partial FIBO Ontology concepts, label, and depth

Label	Depth	Concept	Label	Depth	Concept
C_1	1	Thing	C_{36}	3	NationalIdentificationNumber
C_2	2	Agreement	C_{37}	3	ContractDocument
C_3	2	AutonomousAgent	C_{38}	3	IdentityDocument
C_4	2	Commitment	C_{39}	3	MutualContractualAgreement
C_5	2	ContractualElement	C_{40}	3	ReligiousCorporation
C_6	2	Country	C_{41}	3	BoardAgreement
C_7	2	Date	C_{42}	3	RegistrationIdentifier
C_8	2	Facility	C_{43}	3	ContractParty
C_9	2	FormalOrganization	C_{44}	3	ContractThirdParty
C_{10}	2	Identifier	C_{45}	3	JointStockCompany
C_{11}	2	Jurisdiction	C_{46}	3	PrivatelyHeldCompany
C_{12}	2	LegalDocument	C_{47}	3	PubliclyHeldCompany
C_{13}	2	LegallyCapablePerson	C_{48}	4	Market
C_{14}	2	MonetaryAmount	C_{49}	4	UnilateralContract
C_{15}	2	MutualAgreement	C_{50}	4	VerbalContract
C_{16}	2	NotForProfitCorporation	C_{51}	4	WrittenContract
C_{17}	2	OrganizationCoveringAgreement	C_{52}	4	Adult
C_{18}	2	OrganizationIdentifier	C_{53}	4	Minor
C_{19}	2	PartyInRole	C_{54}	4	NonBindingTermsSet
C_{20}	2	PhysicalAddress	C_{55}	4	NonBindingTerms
C_{21}	2	PhysicalLocation	C_{56}	4	MultilateralTradingFacility
C_{22}	2	StockCorporation	C_{57}	4	OrganizedTradingFacility
C_{23}	2	TransferableContract	C_{58}	4	RegulatedMarket
C_{24}	2	Venue	C_{59}	4	BirthCertificate
C_{25}	2	XMLSchema#string	C_{60}	4	DriversLicense
C_{26}	2	text	C_{61}	4	Passport
C_{27}	2	yesOrNo	C_{62}	4	ContractCounterparty
C_{28}	3	Contract	C_{63}	4	ContractPrincipal

C_{29}	3	Person	C_{64}	5	PromissoryNote
C_{30}	3	ContractTermsSet	C_{65}	5	IncapacitatedAdult
C_{31}	3	ContractualCommitment	C_{66}	5	LegallyCapableAdult
C_{32}	3	ConditionsPrecedent	C_{67}	5	EmancipatedMinor
C_{33}	3	ContractualDefinition	C_{68}	5	RegisteredMultilateralTrading Facility
C_{34}	3	Exchange	C_{69}	5	TransferableContractHolder
C_{35}	3	MarketIdentifier	C_{70}	5	ContractOriginator

4.1.2 Sport Dataset and Ontology

After we found a leaked dataset and ontology in the business domain, it is very important to find another leaked dataset and ontology in a different domain to ensure that the proposed model is applicable and effective on other domain of interest. We evaluate our model in the sports domain using sport ontology and football dataset, which consists of BBC football news and Football leaks dataset [46] [47].

4.1.2.1 *BBC football news dataset*

BBC sport news dataset consists of BBC sport news articles from 2004-2005. In addition, BBC sport news dataset is available publicly online in textual format for research purposes and covers 5 main sports fields including athletics, cricket, football, rugby, and tennis [46]. In our research, we used BBC football news dataset, which consists of 265 BBC football news articles

and we consider them as non-sensitive files. Figure 4.4 shows a sample from the BBC football news dataset.

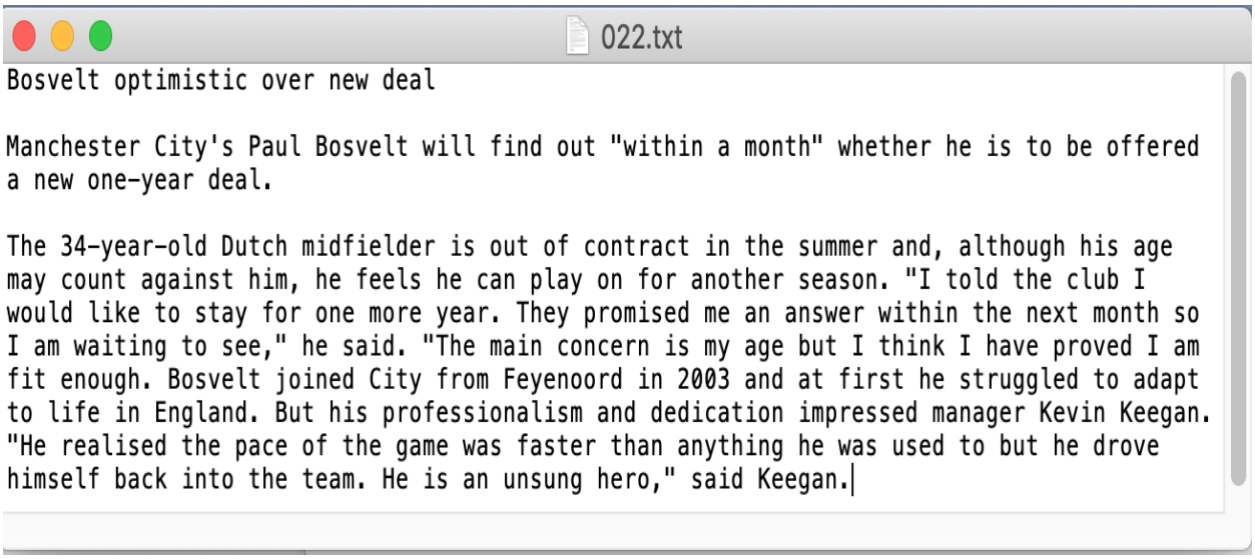


Figure 4.4 A sample from the BBC football news dataset

4.1.2.2 Football leaks dataset

Football leaks dataset is based on one of the largest leaks in sports. In September 2015, football leaks occurred online on a website by Rui Pinto, who at the beginning went by the pseudonym, John. Later, it was found out that Rui Pinto was the actual person behind leaking sensitive football information about football clubs and famous footballers. Leaked data consisted of approximately 18.6 million documents such as emails, contracts, and spreadsheets [48][48]. Der Spiegel with the network *European Investigative Collaborations (EIC)* and its partners used these sensitive documents for investigation. This involved investigative reporting of confidential information by 80 journalists from 12 European media in 11 languages [48][49]. The football leaks are considered a corruption of the European football industry while Rui Pinto declared that he just revealed illegitimate practices that happened within the football world [48].

In 2018, Buschmann and Wulzinger published a book on football leaks that uncovers the dirty deals behind football. The book is structured into several short stories each related to different deals. We created a football leak dataset by saving the different stories in separate files. The football leak dataset consists of the collection of files generated from the aforementioned process. Figure 4.5 shows a sample text from the football leaks dataset.

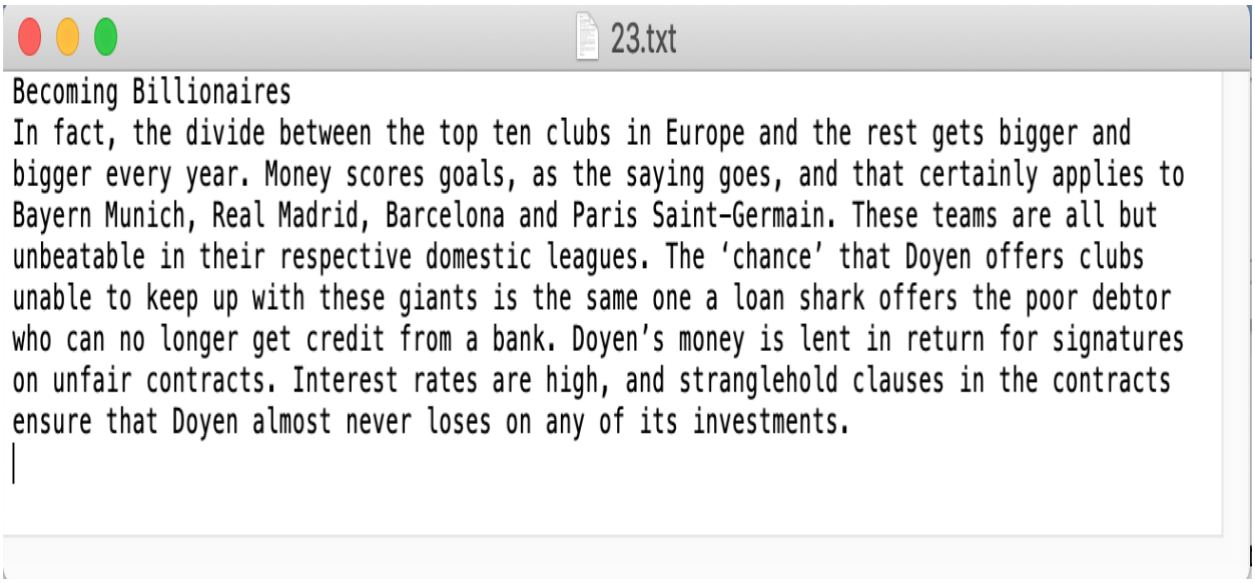


Figure 4.5 A sample text from the football leaks dataset

The football leaks dataset has in total 203 files considered as sensitive files, which were grouped into two subsets based on the first page of a topic, second and third page and so on. To illustrate, the first subset includes the first page of each leaked topic while the second subset has the remaining pages of each topic. Our goal is to use the first subset as the reference sensitive data, while the second subset is used for testing. So in practice, the first subset will represent the protected data, while the second one will be used to test potential leak.

Figure 4.6 shows the statistics of the football dataset including football leaks and BBC football news.

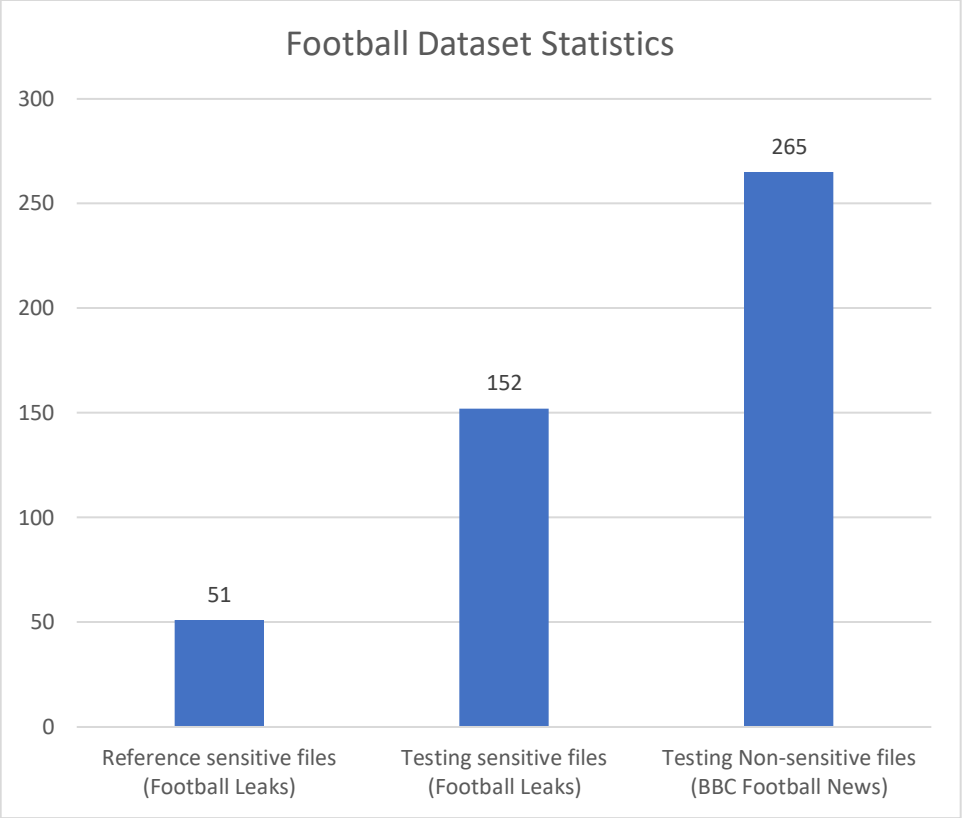


Figure 4.6 Football dataset statistics

4.1.2.3 Sport Ontology

We used Sport ontology, which covers football-related concepts and the relationships between them. Figure 4.7 below shows the concept tree of Sport ontology including concepts, relationships, and concepts' depth [45]. Also, Table 4.2 shows the Sport ontology's concepts, labels, and depths. Furthermore, a sport synonyms keywords' lexicon has been created and provided in our proposed model to assist in finding semantic keywords [53][54].

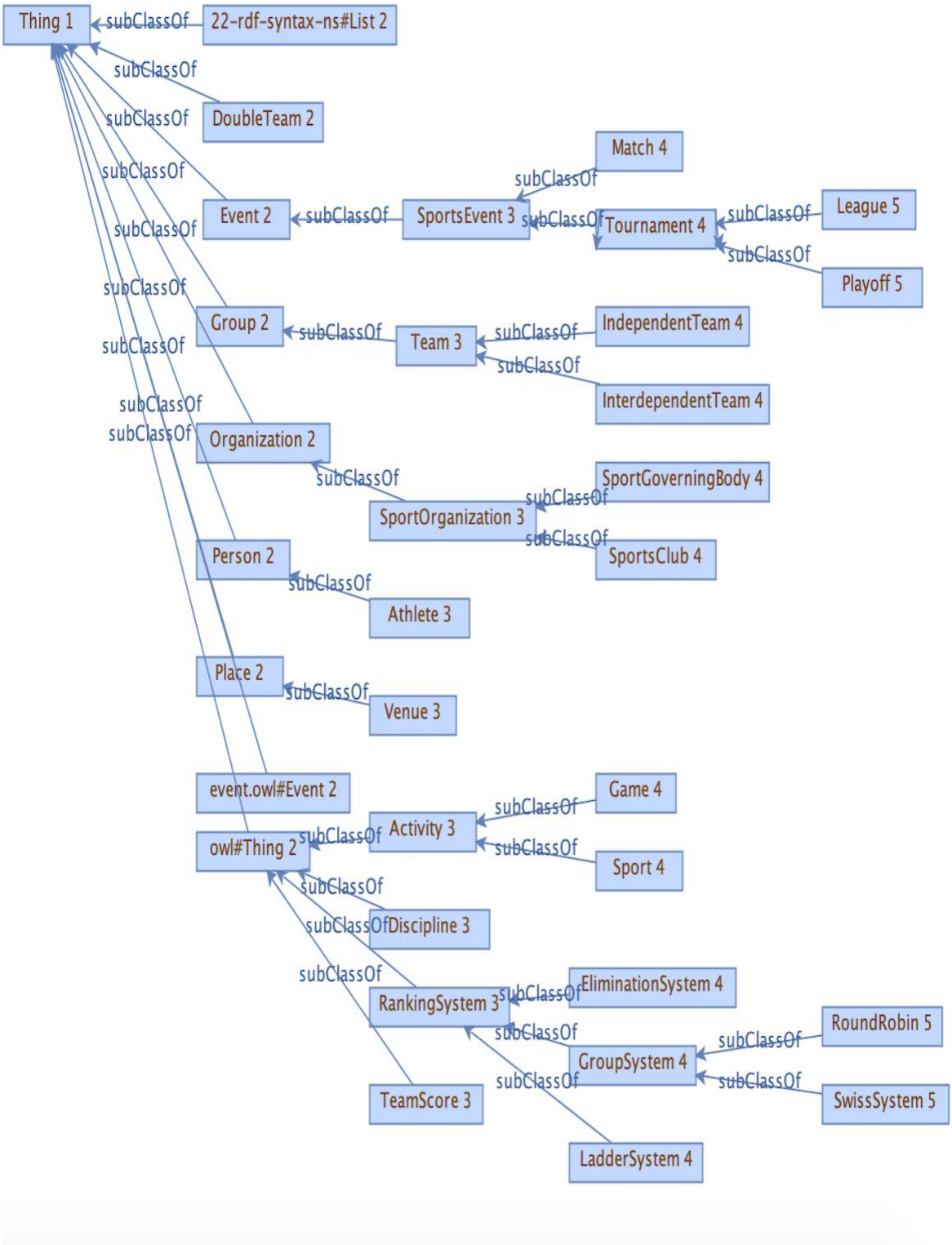


Figure 4.7 The concept tree of Sport ontology

Table 4.2 Sport Ontology concepts, labels, and depths.

Label	Depth	Concept	Label	Depth	Concept
C_1	1	Thing	C_{18}	3	RankingSystem
C_2	2	22-rdf-syntax-ns#List	C_{19}	3	TeamScore
C_3	2	DoubleTeam	C_{20}	4	Match
C_4	2	Event	C_{21}	4	Tournament
C_5	2	Group	C_{22}	4	IndependentTeam
C_6	2	Organization	C_{23}	4	InterdependentTeam
C_7	2	Person	C_{24}	4	SportGoverningBody
C_8	2	Place	C_{25}	4	SportsClub
C_9	2	event.owl#Event	C_{26}	4	Game
C_{10}	2	owl#Thing	C_{27}	4	Sport
C_{11}	3	SportsEvent	C_{28}	4	EliminationSystem
C_{12}	3	Team	C_{29}	4	GroupSystem
C_{13}	3	SportOrganization	C_{30}	4	LadderSystem
C_{14}	3	Athlete	C_{31}	5	League
C_{15}	3	Venue	C_{32}	5	Playoff
C_{16}	3	Activity	C_{33}	5	RoundRobin
C_{17}	3	Discipline	C_{34}	5	SwissSystem

4.2 Evaluation Approach and Metrics

To assess the performance of our model, we calculate the detection rate (DR) and the false positive rate (FPR). DR measures the ability of the model to detect data leakage, while FPR measures its ability to limit false alarms. A false positive occurs when a non-sensitive document is classified as sensitive. A false negative occurs when a sensitive document is falsely classified as non-sensitive; a true detection is just the opposite of a false negative, where a sensitive document is correctly classified as sensitive. The FPR and DR are defined as follows:

$$FPR = \frac{Number\ of\ non\ sensitive\ documents\ classified\ as\ sensitive}{Total\ number\ of\ non\ sensitive\ documents} \qquad (4\text{-}1)$$

$$DR = \frac{Number\ of\ sensitive\ documents\ classified\ as\ sensitive}{Total\ number\ of\ sensitive\ documents} \quad (4\text{-}2)$$

In the Enron email dataset, we consider the initial email from each thread that has two, three, and four emails as sensitive reference emails dataset **(A dataset)** and the set of remaining emails in those threads are considered as a sensitive testing emails dataset **(B dataset)**. In addition, we consider the set of emails in the threads that have only one email as a non-sensitive testing emails dataset **(C dataset)**.

We conducted the evaluation by applying two-fold cross validation on our model as follows. In the 1st round of two-fold cross validation, we check the similarity of each email in B against all emails in A (used as reference) and calculate the number of emails in B that are flagged as dissimilar as per our model (i.e. false negatives). Also, we check the similarity of each email in C against all emails in A and calculate the number of emails in C that have similarity in A (i.e. false positives) as shown in Figure 4.8.

In the 2nd round of two-fold cross validation, we flip between A and B datasets. Then, we check the similarity of each email in A against all emails in B and calculate the number of emails in A that do not have similarity in B (i.e. false negatives). Also, we check the similarity of each email in C against all emails in B and calculate the number of emails in C that have similarity in B (i.e. false positives) as shown in Figure 4.9.

In the football dataset, we consider the 1st thread of football leaks dataset, which has the first page of all topics, as sensitive reference dataset **(A dataset)**, while the 2nd thread which has the remaining pages of each topic, is considered as sensitive testing dataset **(B dataset)**. Also, we consider BBC football news dataset as non-sensitive testing dataset **(C dataset)**. We did the same with the football dataset and evaluated our model by applying two-fold cross validation to calculate the DR and FPR.

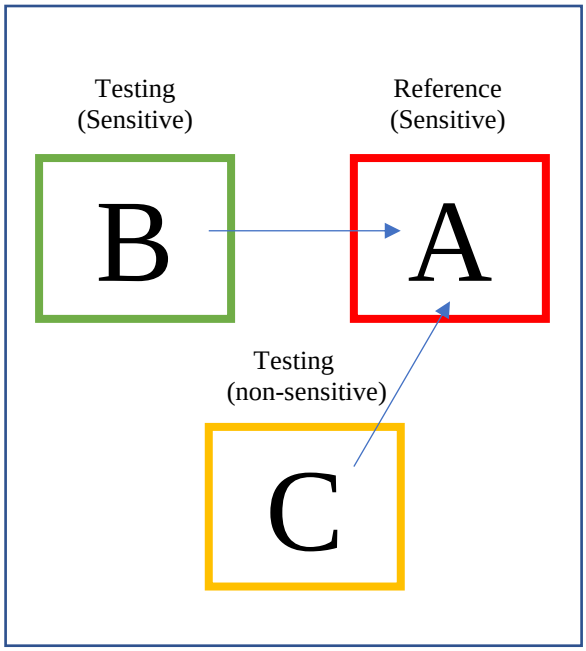


Figure 4.9 1st round of fold cross validation

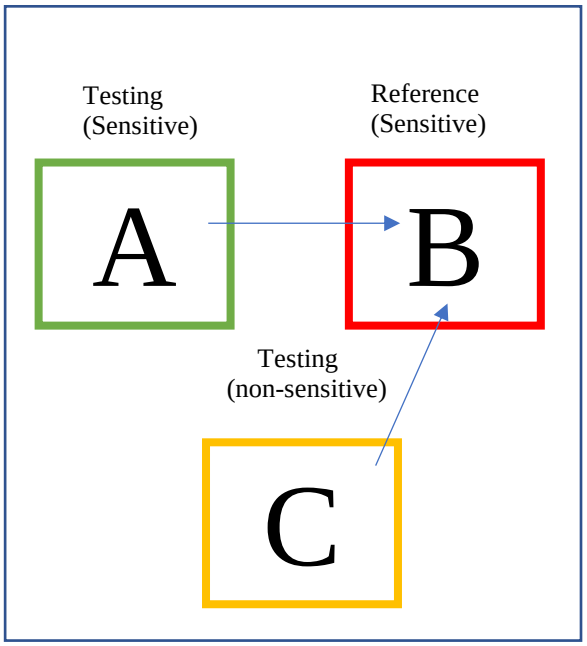


Figure 4.8 2nd round of fold cross validation

4.3 Experimental Evaluation Results

In this section, we describe the experimental evaluation of our proposed model and discuss the obtained results. We conduct various experiments to assess different aspects of the DSS model. Firstly, we apply our proposed model on two different semantic similarity metrics separately, which are the simple frequency and Jaccard index. Next, we evaluate our model by combining the simple frequency and Jaccard metrics with different combination operations on two different datasets of different domain of interests. Then, we use modified Enron emails testing dataset to apply our model. After that, we apply our proposed model on two different domains, which are Business and Sports. Also, we compare our model to two baseline models, namely, TF and TF-IDF models. Next, we apply the combination of semantic similarity metrics (Jaccard and frequency) and semantic relevance metrics on the two datasets. Finally, we implement another semantic similarity metric along with the semantic relevance metric into our model.

4.3.1 Experiment 1

In this experiment, we aim to use two semantic similarity metrics separately in our model and analyze the performance results. We applied our model using the semantic similarity metrics, which are simple frequency (3-7) and Jaccard index (3-8) separately on the Enron email dataset. Our goal is to design an effective detection system, which achieves a high detection rate and a low false positive rate. In fact, a system that exhibits a high false positive rate affects the system negatively by decreasing the confidence in the system and impacting usability. As well, a system that exhibits a low detection rate will miss detecting some sensitive documents. For that in our experiments, it is significant to decide the good detection system, which produces a high detection rate and a low false positive rate. Table 4.3 and Table 4.4 show the FPR and DR results of applying our model using separately the frequency and Jaccard similarity metrics, respectively, by varying the threshold values on the Enron email dataset. (Note: th_f is the threshold of frequency similarity metric and th_j is the threshold of the Jaccard similarity metric)

Table 4.3 Experiment 1: Applying only frequency similarity metric by varying the threshold values on Enron email dataset

	th_f = 55	th_f = 60	th_f = 65	th_f = 70	th_f = 75	th_f = 80	th_f = 85	th_f = 90
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	91.36	91.36	91.36	89	89	89	87.46	86.74
FPR (%)	65.1	59.06	58.22	50.67	44.97	39.1	37.75	35.39

Table 4.4 Experiment 1: Applying only Jaccard similarity metric by varying the threshold values on Enron email dataset.

	th_j = 55	th_j = 60	th_j = 65	th_j = 70	th_j = 75	th_j = 80	th_j = 85	th_j = 90
	Average	Average	Average	Average	Average	Average	Average	Average

DR (%)	87.56	84.47	81.38	80.65	74.28	68.74	63.48	53.77
FPR (%)	11.25	6.71	3.69	3.02	2.35	2.02	1.18	1.18

From these results, we found that applying our model based on only frequency similarity gives higher DR and higher FPR than applying only the Jaccard similarity metric. To illustrate, by applying frequency metric and using threshold th_f =65, we got DR=91.36% and FPR=58.22%, which are suboptimal compared with DR= 81.38% and FPR= 3.69% when applying the Jaccard index and using th_j =65. In fact, this experiment revealed that the performance results of applying Jaccard only are satisfying, getting a high DR and low FPR.

4.3.2 Experiment 2

In this experiment, we aim to study the performance impact of combining two semantic similarity metrics and compare the results with applying the same semantic similarity metrics separately, which are done in experiment 1. This experiment will be done on two different datasets of different knowledge domain.

In this experiment, we run our model using the combination of the two semantic similarity metrics (i.e., frequency and Jaccard metrics), as described in equation (3-7) and (3-8) with different thresholds value on the Enron email dataset. Table 4.5 shows the results of the OR combination between frequency and Jaccard, whereas Table 4.6 shows the results of the AND combination between frequency and Jaccard.

Table 4.5 Experiment 2: Applying OR combination of the frequency with Jaccard for different thresholds on Enron email dataset

	$th_f = th_j = 55$	$th_f = th_j = 60$	$th_f = th_j = 65$	$th_f = th_j = 70$	$th_f = th_j = 75$	$th_f = th_j = 80$	$th_f = th_j = 85$	$th_f = th_j = 90$

	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	91.36	91.36	91.36	89	89	89	87.46	86.74
FPR (%)	65.1	59.06	58.22	50.67	44.97	39.1	37.75	35.39

Table 4.6 Experiment 2: Applying AND combination of the frequency with Jaccard for different thresholds on Enron email dataset.

	$th_f = th_j = 55$	$th_f = th_j = 60$	$th_f = th_j = 65$	$th_f = th_j = 70$	$th_f = th_j = 75$	$th_f = th_j = 80$	$th_f = th_j = 85$	$th_f = th_j = 90$
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	87.56	84.47	81.38	79.93	73.56	65.74	58.31	50.72
FPR (%)	11.25	6.71	3.69	3.02	2.35	1.68	1.18	1.18

The configuration yielding the best performance from Table 4.5 and Table 4.6 is with AND combination of frequency and Jaccard semantic similarity metrics by setting the threshold values for both similarity metrics to 60; this gives DR=84.47% and FPR= 6.71% as shown in Table 4.6.

It emerges from experiment 1 and experiment 2 results that applying the OR combination model and separate models give a higher DR and FPR than the AND combination. The results for the AND combination depicted in Table 4.6 are very encouraging in terms of their high DR and low FPR.

Next, we applied the two different combinations of similarity metrics (frequency and Jaccard) on another domain of interests, which is the sports domain, particularly on football datasets and sport ontology. The performance results of applying OR and AND combinations between frequency and Jaccard on football datasets are shown in Table 4.7 and Table 4.8, respectively.

Table 4.7 Experiment 2: Applying OR combination of the frequency with Jaccard for different thresholds on football datasets.

	$th_f = th_j = 55$	$th_f = th_j = 60$	$th_f = th_j = 65$	$th_f = th_j = 70$	$th_f = th_j = 75$	$th_f = th_j = 80$	$th_f = th_j = 85$	$th_f = th_j = 90$
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	90.53	89.22	67.84	67.84	67.84	67.84	67.84	67.84
FPR (%)	87.04	85.66	66.04	66.04	66.04	65.28	65.28	65.09

Table 4.8 Experiment 2: Applying AND combination of the frequency with Jaccard for different thresholds on football datasets.

	$th_f = th_j = 55$	$th_f = th_j = 60$	$th_f = th_j = 65$	$th_f = th_j = 70$	$th_f = th_j = 75$	$th_f = th_j = 80$	$th_f = th_j = 85$	$th_f = th_j = 90$
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	88.35	85.23	77.35	74.055	67.825	61.26	51.09	40.95
FPR (%)	78.37	70.57	59.63	51.89	43.585	35.61	27.74	21.7

From the two tables above, we notice that applying OR or AND combination between similarity metrics (frequency and Jaccard) on football datasets gives very high DR and FPR. However, applying AND combination gives slightly lower DR and FPR than applying OR combination as shown in Table 4.7 and Table 4.8.

The performance results of this experiment indicate two significant findings. First, we note that applying AND combination of Jaccard and frequency similarity metrics produces better results in terms of high average DR and low average FPR than applying OR combination in the Enron email dataset. Second, the obtained results from applying AND combination of Jaccard and frequency similarity metrics on the Enron email dataset is remarkable but not on football datasets. From this analysis, we can infer that applying AND combination of Jaccard and frequency similarity metrics could generate satisfying performance results in one dataset but not in other ones. As a result of that, we will combine the two semantic similarity metrics (frequency and Jaccard) with the semantic relevance metric to see the impact of their integration on both of the business and sports domain as it will be shown in details in experiment 5.

4.3.3 Experiment 3

In this experiment, we aim to detect any attempts at evading detection by rewriting or modifying the content but keeping the same meaning. For that, we want to assess the effectiveness of our proposed semantic-based detection model when faced with a modified content.

Specifically, we paraphrase 50% of sensitive testing emails of the Enron email dataset by using two online paraphrasing tools along with a human editor to generate a modified version of the original Enron emails [50][51]. Figure 4.10 and Figure 4.11 below present a sample of sensitive testing email before and after paraphrasing, respectively.

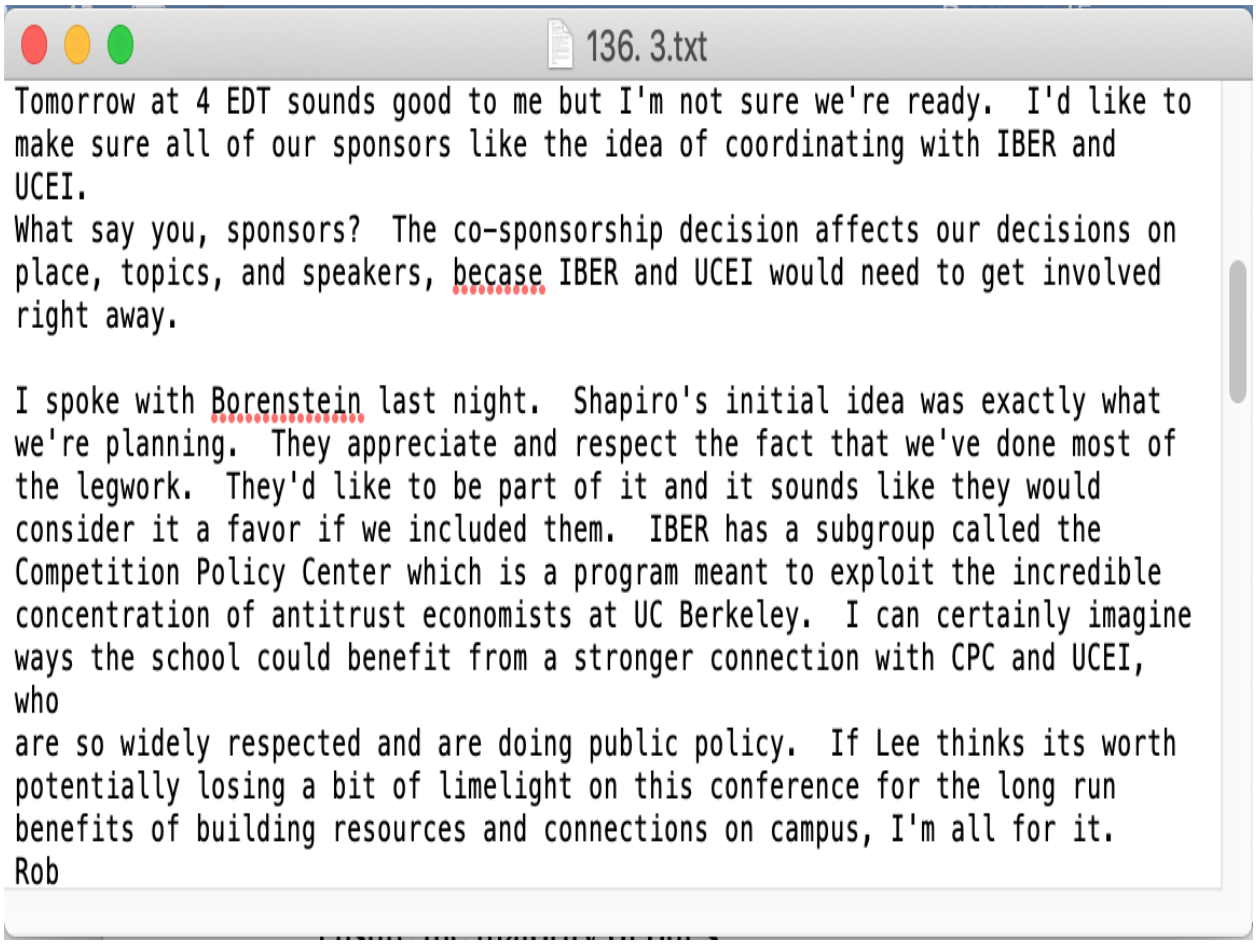


Figure 4.10 A sample of sensitive testing Enron email

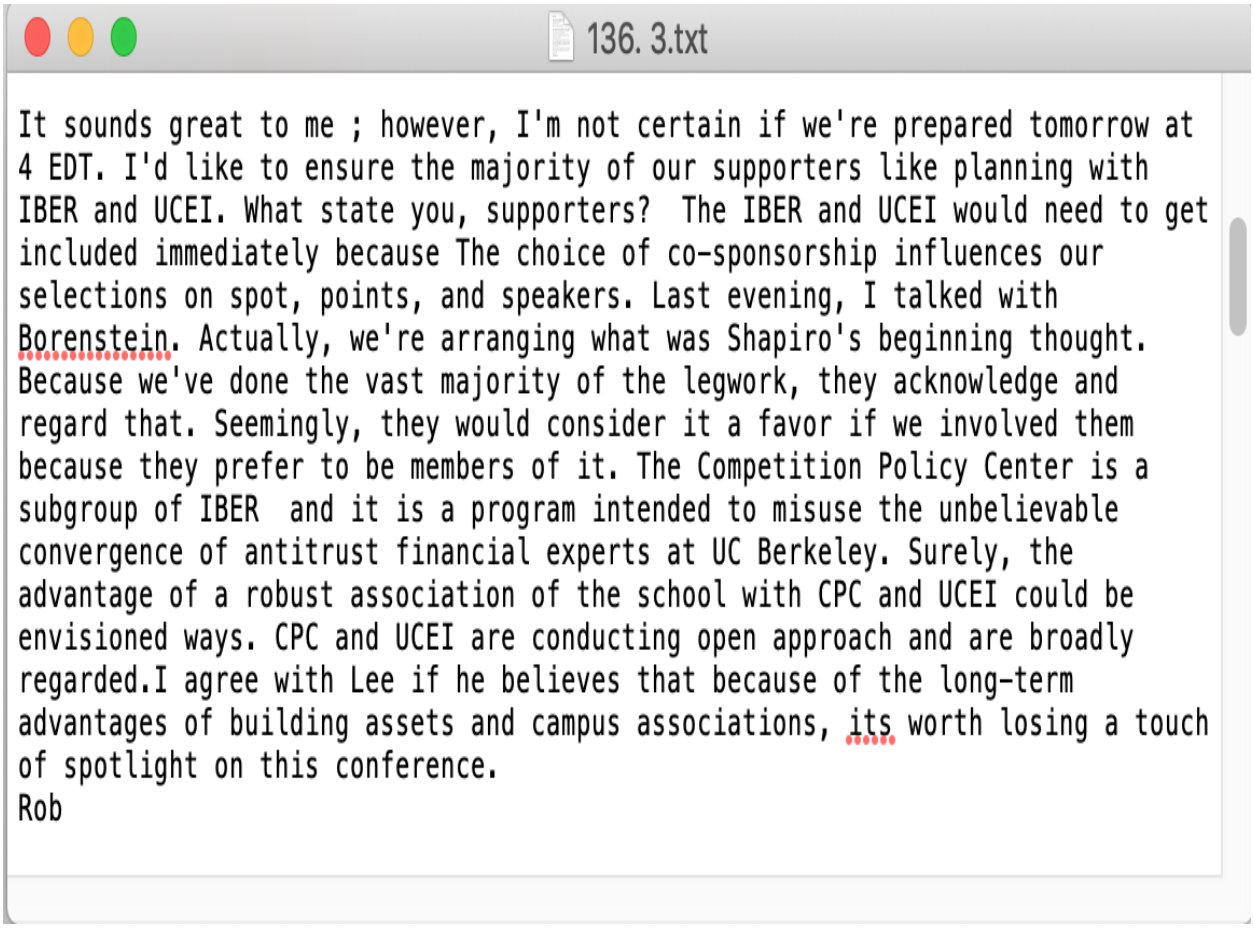


Figure 4.11 A modified version of the testing email shown in Figure 4.10

We ran the model and applied AND combination of frequency and Jaccard similarity metrics on the Enron email dataset containing the modified emails.

Table 4.9 shows the performance results obtained by applying different threshold values.

Table 4.9 Experiment 3: Performance results obtained for different thresholds after paraphrasing 50% of Enron of sensitive testing emails.

	$th_f = th_j = 55$	$th_f = th_j = 60$	$th_f = th_j = 65$	$th_f = th_j = 70$	$th_f = th_j = 75$	$th_f = th_j = 80$	$th_f = th_j = 85$	$th_f = th_j = 90$
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	86.91	82.2	78.525	72.495	66.865	56.945	51.02	37.025
FPR (%)	9.74	5.2	1.34	1.005	0.335	0	0	0

The configuration yielding the best performance from experiment 2 is by setting the threshold for both similarity metrics to 60; this gives DR=84.47% and FPR= 6.71% as shown in Table 4.6. Under this configuration (threshold = 60), we obtain for experiment 3 DR=82.2% and FPR= 5.2% as shown in Table 4.9.

Overall, applying our model on the paraphrased text dataset achieves almost the same results as the original dataset. This underscores the strength of the proposed approach in detecting effectively data leak when the content is modified while keeping the semantic unchanged.

4.3.4 Experiment 4

In this experiment, we aim to compare our model against two different baseline models commonly used in information retrieval separately, that is, term frequency (TF), and term frequency and inverse document frequency (TF-IDF).

TF is the occurrence count of a term in a document divided by the number of concepts in the document. TF-IDF helps even out the effects of too many or few frequently occurring terms.

Consider a set of sensitive documents $M = (d_1, \dots, d_x)$ that must be protected. The TF of the node (or concept) c_j from the ontology concept tree is defined as follows:

$$tf(c_j)_{d_i} = \frac{n(c_j)_{d_i}}{|d_i|} \qquad (4\text{-}3)$$

where $|d_i|$ denotes the number of concepts involved in the document d_i, and $n(c_j)$ is the number of times concept c_j occurs in d_i. Using the same notation as above, *m* corresponds to the total number of concepts in the ontology concept tree, and $1 \leq j \leq m$.

The inverse document frequency (IDF) for c_j with respect to *M* is computed as follows:

$$idf(c_j)_M = 1 + log\left(\frac{x}{\{d:M|c_j \in d\}}\right) \quad (4\text{-}4)$$

When concept c_j does not appear in any documents, then the denominator of $idf(c_j)_M = 1 + log\left(\frac{x}{\{d:M|c_j \in d\}}\right)$ (4-4) will be equal to zero, which in turn gives an infinite value for $idf(c_j)$. To avoid this, we assign -1 to $idf(c_j)$ if c_j does not appear in any document.

The TF-IDF for c_j with respect to *M* is computed as follows:

$$tf{-}idf(c_j)_{d_i,M} = tf(c_j)_{d_i} \times idf(c_j)_M \quad (4\text{-}5)$$

The TF matrix for document d_i is defined as follows:

$$TF(d_i) = \left[tf(c_j)_{d_i}\right]_{1 \le j \le m} \quad (4\text{-}6)$$

Similarly, the TF-IDF matrix for document d_i is defined as:

$$TF - IDF(d_i) = \left[tf{-}idf(c_j)_{d_i,M}\right]_{1 \le j \le m} \quad (4\text{-}7)$$

Given a document *d* being checked for data leaks against reference M, the CS is applied against each of the sensitive documents $d_i \in M$. This gives for TF,

$$CS(d, d_i) = \frac{\sum_{j=1}^{m} tf(c_j)_d \cdot tf(c_j)_{d_i}}{\sqrt{\sum_{j=1}^{m}\left(tf(c_j)_d\right)^2} \times \sqrt{\sum_{j=1}^{m}\left(tf(c_j)_{d_i}\right)^2}} \quad (4\text{-}8)$$

whereas for TF-IDF,

$$CS(d, d_i) = \frac{\sum_{j=1}^{m}(tf{-}idf(c_j)_{d,M} \times tf{-}idf(c_j)_{d_i,M})}{\sqrt{\sum_{j=1}^{m}(tf{-}idf(c_j)_{d,M})^2} \times \sqrt{\sum_{j=1}^{m}\left(tf{-}idf(c_j)_{d_i,M}\right)^2}} \quad (4\text{-}9)$$

In this experiment, we apply the TF and TF-IDF baseline models separately according to the different threshold values of the cosine similarity of the TF vectors (Thr_1) and TF-IDF vectors (Thr_2). The TF and TF-IDF vectors are created based on the FIBO ontology concepts.

In the TF baseline model, we calculate the cosine similarity between the TF vectors of the monitored document *d* against each sensitive documents $d_i \in M$. The cosine similarity value is compared against a threshold to make a decision regarding the similarity of the documents. In other words, if the cosine similarity value between the monitored document *d* and at least one sensitive document $CS(d, d_i)$ is above a certain threshold (Thr_1), then we will consider document *d* to be similar to d_i; otherwise, they are not similar.

We apply the same steps for the TF-IDF baseline model by taking into account the TF-IDF vectors and TF-IDF threshold (Thr_2).

In addition, we apply baseline models using Enron email dataset and FIBO ontology. Also, we calculate the average DR and FPR based on two-fold cross validation for both baseline models. Table 4.10 and Table 4.11 show the results of applying the TF and TF-IDF baseline models, respectively.

Table 4.10 Experiment 4: TF baseline model for different thresholds

	Thr_1=55	Thr_1=60	Thr_1=65	Thr_1=70	Thr_1=75	Thr_1=80	Thr_1=85	Thr_1=90
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	76.02	75.20	69.84	63.47	50.39	41.23	32.87	25.97
FPR (%)	67.96	62.76	59.06	53.36	39.60	31.54	25.34	21.82

Table 4.11 Experiment 4: TF-IDF baseline model for different thresholds

	Thr_2=55	Thr_2=60	Thr_2=65	Thr_2=70	Thr_2=75	Thr_2=80	Thr_2=85	Thr_2=90
	Average	Average	Average	Average	Average	Average	Average	Average
DR (%)	77.66	70.75	64.49	53.58	43.59	37.51	31.42	27.42
FPR (%)	64.77	55.71	50.68	42.96	37.75	32.72	27.52	21.98

The configuration yielding the best performance in Experiment 2 is when we applied our model based on AND combination of both frequency and Jaccard similarity metrics. As a result, the

best results using 60 for both threshold values Thr_f and Thr_j was DR=84.465% and FPR=6.71%, which was very encouraging. On the other hand, in experiment 4 under such condition (Thr_1 = 60, $Thr_2=60$) in Table 4.10 and

Table 4.11, we obtain DR=75.20% , FPR=62.76% and DR=70.75% , FPR=55.71% for TF and TF_IDF baseline models, respectively. From the aforementioned results, the TF and TF-IDF baseline models give both a high DR and high FPR. This finding indicates that our semantic signature model outperforms the above baseline models. However, we intend to further explore more experimental configurations to improve on these results.

4.3.5 Experiment 5

In this experiment, we aim to study the detection ability of our proposed model using a combination of semantic similarity metrics (frequency (3-7) and Jaccard (3-8)) and semantic relevance metric (3-14) to measure the semantic overlap between documents and to detect suspicious documents in two different datasets.

For that, we applied our model on the Enron email dataset and four modules of FIBO ontology, which are Corporation, Contract, Market, and people, based on predefined thresholds of frequency, Jaccard, and average relevancy set to 60, 60, and 0.35, respectively. We calculate the average DR and FPR based on two-fold cross validation. (Note: th_f is the threshold of frequency similarity metrics, th_j is the threshold of the Jaccard similarity metrics, and th_{rel} is the threshold of the average semantic relevancy). Based on experiment 2's interesting finding, we consider the AND combination between frequency and Jaccard semantic similarity metrics to be applied in this experiment. This experiment is branched into two sections as follows:

1- Business domain by using Enron email dataset and FIBO ontology as its results will be shown in Table 4.12 and Table 4.13.
2- Sports domain by using football datasets and sport ontology as its results will be shown in Table 4.14 and Table 4.15.

Then, each section of this experiment involves two sub-experiments based on two different combinations between semantic similarity and semantic relevance metrics as follows:

1- (Simple frequency (3-7) AND Jaccard Index (3-8)) OR average relevance (3-14).
2- (Simple frequency (3-7) AND Jaccard Index (3-8)) AND average relevance (3-14).

This experiment applied the OR combination between semantic similarity metrics (frequency and Jaccard) with semantic relevancy metric (average relevancy). Table 4.12 shows the average of DR and FPR of two-fold cross validation of applying OR combination on Enron email dataset.

Table 4.12 Experiment 5: Performance results obtained based on predefined thresholds for combined semantic similarity and semantic relevance metrics with OR combination on Enron email dataset.

Ontology Modules	(th_f=60 **AND** th_J=60) **OR** th_{rel}=0.35	
People, Corporations, Markets, Contracts		Average
	DR (%)	87.555
	FPR (%)	81.54

Based on predefined thresholds of frequency=60, Jaccard=60, and average relevance=0.35, the results show high average DR and FPR, which are 87.56% and 81.54%, respectively. The average FPR is disappointing since it is so high. For that, we will apply a different combination between semantic similarity metrics and semantic relevance metric to get better performance results.

Next, we applied AND combination between semantic similarity metrics (frequency and Jaccard) with semantic relevancy metric (average relevancy). Table 4.13 shows the average of DR and FPR of two-fold cross validation of applying AND combination on Enron email dataset.

Table 4.13 Experiment 5: Performance results obtained based on predefined thresholds for combined semantic similarity and semantic relevance metrics with AND combination on Enron email dataset.

Ontology Modules	(th_f=60 **AND** th_J=60) **AND** th_{rel}=0.35	
People, Corporations, Markets, Contracts		Average
	DR (%)	77.56
	FPR (%)	0

The results show relatively high average DR, which is 77.56% and low average FPR, which is 0%. Under the same configuration of thresholds (frequency = 60, Jaccard=60, and average relevancy= 0.35), we found that OR combination gives higher DR and FPR than AND combination. Obviously, we achieve better results in AND combination in Table 4.13 than OR combination in Table 4.12 in terms of detection effectiveness represented by high DR and low FPR. Also based on AND combination, when we compare the results of this experiment with experiment 2, we note an interesting finding that experiment 5 got a slightly lower DR and FPR than experiment 2. To illustrate, experiment 5 achieves a considerable difference in FPR, which dropped from 6.38 to 0%. From the first section of this experiment, we can say that combining semantic relevance metric to semantic similarity metrics improves the performance results of the detector.

As well with football datasets and sport ontology, we apply OR and AND combination of semantic similarity metrics and semantic relevance metric based on predefined thresholds of frequency, Jaccard, and average relevancy, which are set to 60, 60, and 0.35, respectively. The

obtained performance results of applying OR and AND combination on football datasets are shown in Table 4.14 and Table 4.15, respectively.

Table 4.14 Experiment 5: Performance results obtained based on predefined thresholds for combined semantic similarity and semantic relevance metrics with OR combination on football datasets.

Ontology Modules	(th_f=60 **AND** th_J=60) **OR** th_{rel}=0.35	
Sport		Average
	DR (%)	90.48
	FPR (%)	81.135

Table 4.15 Experiment 5: Performance results obtained based on predefined thresholds for combined semantic similarity and semantic relevance metrics with AND combination on football datasets.

Ontology Modules	(th_f=60 **AND** th_J=60) **AND** th_{rel}=0.35	
Sport		Average
	DR (%)	84.98
	FPR (%)	19.7

In the two tables above, we can see that OR combination produces very high average DR and FPR compared to AND combination results. Also in Table 4.15, AND combination generates a high DR and a relatively low FPR, which in fact needs to be enhanced to detect similar texts semantically.

From this experiment and all the previous experiments, we have a clear insight that OR combination affects the performance results negatively by generating a very high FPR, which is unacceptable for any detection system. In addition, using football datasets, we found notable differences between experiment 2 when we applied a combined semantic similarity metrics (frequency and Jaccard) as shown in Table 4.8 and when we applied integrated semantic similarity metrics with the semantic relevancy metric as shown in Table 4.15. To illustrate, under the same configuration of thresholds, which sets both frequency and Jaccard to 60 and relevancy

average to 0.35, in Table 4.8 we obtain DR= 85.23% and FPR=70.57%, while in Table 4.15 when we integrate semantic similarity with semantic relevance, we obtain DR= 84.98% and FPR=19.7%. We observe a sharp decline in FPR between Table 4.8 and Table 4.15, which confirms that detecting semantics between documents is more effective when we integrate semantic similarity with semantic relevance metrics.

Obviously with the football dataset, since we got 19.7 % for FPR as shown in Table 4.15 which is unsatisfactory compared to the Enron email dataset for which 0% was obtained for FPR as shown in table 4.13, we decided to run another experiment to combine another semantic similarity metric to the same semantic relevance metric. This step will be evaluated to assess the model and indicate the difference between applying semantic relevance metric combined with two different semantic similarity metrics. In experiment 6, we will apply that, aiming to enhance the performance results of the detection model.

4.3.6 Experiment 6

In this experiment, we aim to evaluate our model on another semantic similarity metric (3-10) combined with the semantic relevance metric (3-14) to measure the semantic overlap between documents and to detect suspicious documents. We conducted this experiment on two different domains of interests, which are the business and sports domain, one at a time.

We calculate the average DR and FPR based on two-fold cross validation. (Note: th_{sim} is the threshold of semantic similarity while th_{rel} is the threshold of relevancy metric)

Table 4.16 shows the performance results obtained by applying the combination of semantic similarity (3-10) and semantic relevance metrics (3-14) using the Enron email dataset and FIBO ontology.

Table 4.16 Experiment 6: Performance results obtained by applying the combination of semantic similarity and relevancy metrics on Enron email dataset.

		$thr_{sim}=0.45$ AND $thr_{rel}=0.30$
Ontology Modules		Average
People, Corporations, Markets, Contracts	DR (%)	92.085
	FPR (%)	15.27

The best results obtained using 0.45 for similarity threshold and 0.30 for relevancy threshold values were DR=92.085% and FPR=15.27%.

Next, Table 4.17 shows the performance results obtained by applying the combination of semantic similarity (3-10) and semantic relevancy metrics (3-14) using the football dataset and sport ontology.

Table 4.17 Experiment 6: Performance results obtained by applying the combination of semantic similarity and relevancy metrics on Football dataset.

		$thr_{sim}=0.65$ AND $thr_{rel}=0.5$
Ontology Modules		Average
Sport	DR (%)	72.79
	FPR (%)	3.4

The best results obtained using 0.65 for similarity threshold and 0.50 for relevancy threshold values were DR=72.79% and FPR=3.4%.

Table 4.16 and Table 4.17 show that we achieve satisfying performance by obtaining high DR and low FPR results of combining semantic similarity and semantic relevance metrics on both datasets of different knowledge. We note a crucial finding that applying semantic similarity metric (3-10) combined with semantic relevance metric (3-14) in experiment 6 is more adequate than applying semantic similarity metrics (frequency (3-7) and Jaccard (3-8)) with semantic relevance metric (3-14) in experiment 5. That was shown clearly in experiment 6's satisfying results with both datasets while experiment 5's results were satisfying just with the Enron email dataset. In other

words, semantic similarity metric (3-10) shows its wider applicability and higher efficiency on different datasets than semantic similarity metrics (frequency (3-7) and Jaccard (3-8)).

4.4 Summary

This chapter presents and discusses the results of several experiments. We draw several findings from the results of those experiments as following. Our semantic similarity model is effective in terms of detection and its applicability on the different domains of knowledge, which is clearly shown by applying our model on business and sports dataset. Also, the comparison of our model to baseline models such as TF and TF-IDF models, show that it outperforms these baseline models. As well, our semantic model achieved satisfying detection results when it is evaluated on a modified dataset; this shows its robustness in countering evasion tactics based on content rewriting. Moreover, it can be noted that combining semantic similarity metric to semantic relevance metric gives better results in detecting semantic commonalities between any suspicious and sensitive documents.

Chapter 5

Conclusion

One of the main threats faced by any organization that maintains sensitive digital assets is the threat posed by malicious insiders. A malicious insider is an authorized member of the organization (e.g. a disgruntled employee) that intentionally and negatively affects the integrity, confidentiality, or availability of the organization's resources including data, system, or network [52]. Since the sensitive information in any organization is vulnerable to be leaked, damaged, or lost, it is necessary to secure this information along with assigning desired privilege to each employee. Some of these insiders, whose aim is to threaten the organization security, leverage their privileges to leak sensitive data. Some of them can do that smartly by altering sensitive file's content but keeping the same meaning in order to remove any suspicion around using or transferring that file. Existing Data loss prevention (DLP) mechanisms are inefficient when confronted with this kind of data alteration.

The purpose of our proposed research is to address the aforementioned challenge by developing a new DLP approach to monitor transmitted data by checking their content semantically. From previous works, there is a clear limitation in searching and comparing contents semantically.

Our proposed model extracts a summarized form of the semantic of each document that we refer to as the document semantic signature. By comparing the similarity of the signature of a monitored document against reference signatures for sensitive documents (i.e. to be protected), we are able to detect effectively potential data leakage, even when the content is altered while keeping the semantic unchanged. Basic components of our proposed approach have been

defined, developed, and evaluated against two existing public leaked datasets, yielding very encouraging performance results. Several practical issues identified above must still be addressed.

5.1 Contribution Summary

We have defined a new model to capture data leakage by tracking the content semantically. The proposed model allows expressing and extracting the semantic from the document content. The extracted semantic is summarized under the form of the document semantic signature (DSS). We have defined, using a combination of available similarity metrics, an approach to match semantic signatures and detect potential data leakage. The proposed model has been evaluated experimentally using two different datasets covering different knowledge domains, namely a subset of the Enron dataset and the Football news dataset, yielding very performance results.

Our signature matching algorithm uses, in combination, two different similarity metrics. Also in the experiments, we have studied each of the similarity metrics separately, and compare them to the combined model.

Our proposed semantic signature model depends on the dimension of the ontology. A typical ontology may involve thousands of concepts, so the signature size can be very large. We have studied the time and space complexity of the matching algorithm and they are acceptable.

Document content semantic comprises both a set of concepts and a set of relationships among the concepts. Our current semantic signature model focuses essentially on the concepts and important semantic information conveyed by the conceptual relationships. We have extended the model and taken into account the semantic relation between concepts, both explicit and

implicit relations. This involved using a logical framework such as the Ontology Description Logics (DL) to analyze the relationships and infer hidden ones.

We ran our model using another semantic similarity metric combined with semantic relevance metric and have gotten encouraging performance results. More focused evaluation was conducted to study the impact of data modification on the model and ensure the effectiveness of our proposed model. Also, we have compared our model to baseline models, such as models based on computing and tracking TF-IDF measures from documents, and found that our model outperforms these previous models.

5.2 Future Work

In our future work, we plan to study our model on a wider variety of domain of interests such as medical, technical, and industrial domains. In addition, finding leaked textual datasets in any domain of knowledge is challenging. Still, we will look for other leaked datasets that belong to other domains, such as any of the aforementioned domains, to run our model on several domains. More extensive analysis of the data will also be conducted by considering different scenarios.

We would like to extend our model to provide a full suite DLP, which includes detection and prevention capabilities to protect sensitive data in any organization. Also, we will enhance our models' capability by addressing data in any of the three information states. For that, we intend to incorporate some prevention methods, such as block the document, label or notify the user, and report the administrator, to secure critical data inside organizations' network.

Moreover, we will aim to study the integration of machine learning techniques to our model. There are obvious limitations in using machine learning techniques for DLP as discussed

in the related chapter (i.e. chapter 2). However, selective integration of machine learning by using it for specific tasks or aspects could yield improved detection capability.

Furthermore, we will build an ontology for a specific domain, which helps to construct concepts and their relations based on the critical data content and the organization's requirements. As well, framing and structuring customized ontology helps to aggregate concepts of two domains under an ontology.

Our current signature model is relative to a specific domain ontology, which captures only knowledge related to the corresponding domain. However, it is not unusual that a sensitive document may carry important information related to multiple different domains, and thereby to different ontologies. We will expand our semantic signature model to cover such a situation where multiple ontologies will be linked to the same signature, rather than having to maintain multiple signatures for the same document.

Bibliography

[1] Grammatikis, P.I.R., Sarigiannidis, P.G. and Moscholios, I.D., "Securing the Internet of Things: Challenges, Threats and Solutions". *Internet of Things*, 2018, pp. 41-69, doi: https://doi.org/10.1016/j.iot.2018.11.003.

[2] Petrakis, E.G., Sotiriadis, S., Soultanopoulos, T., Renta, P.T., Buyya, R., Bessis, N., "Internet of Things as a Service (iTaaS): Challenges and Solutions for Management of Sensor Data on the Cloud and the Fog". *Internet of Things*, *3*, 2018, pp.156-174, doi: https://doi.org/10.1016/j.iot.2018.09.009.

[3] CA Technologies, "Insider threat 2018 report", Available: https://www.ca.com/content/dam/ca/us/files/ebook/insider-threat-report.pdf. Accessed 6 March 2019.

[4] Di Martino, B., Rak, M., Ficco, M., Esposito, A., Maisto, S.A., Nacchia, S., "Internet of things reference architectures, security and interoperability: A survey". *Internet of Things*, *1*, 2018, pp.99-112, doi: https://doi.org/10.1016/j.iot.2018.08.008.

[5] Kowalski, E., Cappelli, D., Moore, A., U.S. Secret Service and CERT/SEI Insider Threat Study: Illicit Cyber Activity in the Information Technology and Telecommunications Sector Carnegie Mellon Software Engineering Institute, Pittsburgh, 2008.

[6] Ramachandran, R., Neelakantan, S. and Bidyarthy, A.S., 2011, December. Behavior model for detecting data exfiltration in network environment. In *2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application* (pp. 1-5). IEEE, 2011.

[7] Costa, D.L., Collins, M.L., Perl, S.J., et al., An Ontology for Insider Threat Indicators Development and Applications (Carnegie-Mellon University, Pittsburgh, Software Engineering Inst, 2014

[8] Alhindi H., Traore I., Woungang I. (2019) Data Loss Prevention Using Document Semantic Signature". In: Woungang I., Dhurandher S. (eds) 2nd International Conference on Wireless Intelligent and Distributed Environment for Communication (WIDECOM 2019), Milan, Italy, 2019, pp 75-99. Lecture Notes on Data Engineering and Communications Technologies, vol 27. Springer. https://doi.org/10.1007/978-3-030-11437-4_7.

[9] Alhindi H., Traore I., Woungang I. "Preventing Data Leak through Semantic Analysis." *Internet of Things:* (2019): 100073.

[10] "Data Loss Prevention: A Holistic approach", DLP Whitepaper, SecureReading. [Online] Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=2ahUKEwiUj_zatPHjAhWeIDQIHW6mBXsQFjADegQIBRAC&url=https%3A%2F%2Fsecurereading.com%2Fwp-content%2Fuploads%2F2016%2F08%2FData-Leakage-Prevention.pdf&usg=AOvVaw0nZTDVLGgS-j4EpZ2VS-Q3 [Accessed July 29, 2019]

[11] "Morgan Stanley Fires Rogue Employee After Customer Data Leak", [Online] Available: https://www.forbes.com/sites/antoinegara/2015/01/05/morgan-stanley-fires-rogue-employee-after-customer-data-leak/#39d2e8b67b7e Marsh data leak, [Accessed July 29, 2019]

[12] Katz, G., Elovici, Y., Shapira, B. (2014). "CoBAn: A context based model for data leakage prevention." *Information sciences* 262 (2014): pp.137-158.

[13] The most infamous data breaches, [Online] Available: https://www.techworld.com/security/uks-most-infamous-data-breaches-3604586/, [Accessed July 29, 2019]

[14] Data Leakage Prevention best practices, [Online] Available: https://www.compuquip.com/blog/8-data-leakage-prevention-best-practices , [Accessed July 29, 2019]

[15] Data Loss Prevention, [Online] Available: https://www.cisco.com/c/en/us/products/security/email-security-appliance/data-loss-prevention-dlp.html, [Accessed July 29, 2019]

[16] Liu, S. and Kuhn, R., 2010. Data loss prevention. *IT professional*, *12*(2), pp.10-13.

[17] Data Loss Prevention, [Online] Available: https://www.imperva.com/learn/data-security/data-loss-prevention-dlp/ [Accessed July 29, 2019]

[18] Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V., (2016). "A survey on data leakage prevention systems". *Journal of Network and Computer Applications*, *62*, pp.137-152.

[19] "Data Loss Prevention: Keeping your sensitive data out of the public domain", Whitepaper, October 2011

[20] "Understanding and Selecting a Data Loss Prevention Solution", Whitepaper, Securosis, L.L.C. [Online] Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwjjzsXJtfHjAhWiCjQIHQYNBxwQFjABegQIBBAC&url=https%3A%2F%2Fsecurosis.com%2Fassets%2Flibrary%2Freports%2FUnderstanding_and_Selecting_DLP.v3_FINAL_.pdf&usg=AOvVaw1X8YdSpVUiFfi3r_eXti90 [Accessed July 29, 2019]

[21] "Data Loss Prevention: moving beyond perimeter security with a flexible, in-depth approach to protecting data", DLP Whitepaper, CDW. [Online] Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&cad=rja&uact=8&ved=2ahUKEwiMyrz5tPHjAhWUFzQIHR_RAPAQFjAJegQIAxAC&url=https%3A%2F%2Fwe

bobjects.cdw.com%2Fwebobjects%2Fmedia%2Fpdf%2FSolutions%2Fsecurity%2FData-Loss-Whitepaper.pdf&usg=AOvVaw02evjzc9zAPj3iyb9WYX9W [Accessed July 29, 2019]

[22] Tahboub, R. and Saleh, Y., 2014, January. Data leakage/loss prevention systems (DLP). In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-6). IEEE, 2014.

[23] What are the consequences of data loss, [Online] Available: https://www.unitrends.com/blog/what-are-the-consequences-of-data-loss, [Accessed July 29, 2019]

[24] Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., Gritzalis, D. (2010). "An insider threat prediction model", In International Conference on Trust, Privacy and Security in Digital Business, Springer, pp. 26-37, 2010. [online]. Available: https://resources.sei.cmu.edu/asset_files/WhitePaper/2008_019_001_52266.pdf; [Accessed: November 22, 2017]

[25] Udoeyop, A. W. (2010). "Cyber profiling for insider threat detection", 2010.

[26] Liu, Y., Corbett, C., Chiang, K., Archibald, R., Mukherjee, B., Ghosal, D. (2009). "SIDD: A framework for detecting sensitive data exfiltration by an insider attack", In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on IEEE, pp. 1-10, 2009.

[27] Ragavan, H. (2012). "Insider threat mitigation models based on thresholds and dependencies", University of Arkansas, 2012.

[28] Raman, P., Kayacık, H. G., Somayaji, A. (2011). "Understanding data leak prevention", In 6th Annual Symposium on Information Assurance (ASIA'11), pp. 27-31, 2011.

[29] Liu, S., and Kuhn, R. (2010). "Data loss prevention", IT professional, vol. 12(2), 2010.

[30] Hart, M., Manadhata, P., Johnson, R. (2011). “Text classification for data loss prevention”, In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 18–37, 2011.

[31] Stamati-Koromina, V., Ilioudis, C., Overill, R., Georgiadis, C. K., Stamatis, D. (2012). “Insider threats in corporate environments: a case study for data leakage prevention”. In Proceedings of the Fifth Balkan Conference in Informatics pp. 271-274. ACM., 2012.

[32] Canbay, Y., Yazici, H., Sagiroglu, S. (2017). “A Turkish language based data leakage prevention system”, In Digital Forensic and Security (ISDFS), 2017 5th International Symposium, pp. 1-6, IEEE., 2017.

[33] Du, D., Lu Y, Richard R. Brooks. "Semantic similarity detection for data leak prevention." *Proceedings of the 10th Annual Cyber and Information Security Research Conference*. ACM, 2015.

[34] Vodithala, S. and Pabboju, S. (2017). "A Keyword Ontology For Retrieval of Software Components," International Journal of Control Theory and Applications, vol. 10, no. 19, pp. 177-182, 2017.

[35] Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E. (2011). “Semantically enhanced information retrieval: An ontology-based approach”, Web semantics: Science, services and agents on the world wide web, vol. 9(4), pp. 434-452, 2011.

[36] Doing-Harris, K., Livnat, Y., Meystre, S. (2015). “Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system”, Journal of biomedical semantics, vol. 6, 2015.

[37] Liu, H. Z., Bao, H., Xu, D. (2012). “Concept vector for similarity measurement based on hierarchical domain structure”, Computing and informatics, vol. 30(5), pp. 881-900, 2012.

[38] Corley, C., & Mihalcea, R. (2005). “Measuring the semantic similarity of texts”. In Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment, Association for Computational Linguistics, pp. 13-18, 2005.

[39] Onix. “Onix Text Retrieval Toolkit API Reference”, [Online] Available: http://www.lextek.com/manuals/onix/stopwords1.html, [Accessed: November 14, 2017]

[40] Enron. "Enron Email Dataset," [Online]. Available: https://www.cs.cmu.edu/~./enron/. [Accessed: October 20, 2017].

[41] Klimt B. and Y. Yang (2004). The Enron Corpus: A new dataset for email classification research. In Machine learning: ECML 2004, pages 217–226. Springer, 2004.

[42] Saad, S., Traore, I. A Semantic-Based Approach for Heterogeneous Multi-Sensor Alerts Aggregation, 2012.

[43] FIBO. Financial Industry Business Ontology, [Online]. Available: https://www.edmcouncil.org/financialbusiness. [Accessed: October 20, 2017].

[44] Business Balls. [Online]. Available: http://www.businessballs.com/business-thesaurus.htm. [Accessed: October 19, 2017].

[45] Sports ontology, [Online]. Available: https://github.com/Tobion/Sports-Ontology/blob/master/ontology.owl. [Accessed: July 5, 2019].

[46] BBC Datasets, [Online]. Available: http://mlg.ucd.ie/datasets/bbc.html. [Accessed: July 5, 2019].

[47] Buschmann, R., and Wulzinger, M. Football leaks: Uncovering the dirty deals behind the beautiful game. Faber & Faber, 2018.

[48] Football Leaks website, [Online]. Available: https://footballleaks2015.wordpress.com/ [Accessed: July 5, 2019].

[49] De Spiegel, [Online]. Available: https://www.spiegel.de/international/world/interview-with-football-leaks-whistleblower-rui-pinto-a-1251121.html, [Accessed: online July 5, 2019].

[50] Project Topics, [Online]. Available: https://paraphrase.projecttopics.org/free-paraphrasing-tool. [Accessed: February 15, 2019].

[51] QuillBot App, [Online]. Available: https://quillbot.com/app. [Accessed: February 15, 2019].

[52] Mahajan, A., Sharma, S., "The Malicious Insiders Threat in the Cloud." International Journal of Engineering Research and General Science vol. 3.2, pp. 245-256, 2015.

[53] Thesaurus, [Online] Available: https://www.thesaurus.com/browse/ , [Accessed: July 5, 2019].

[54] Power Thesaurus, [Online] Available: https://www.powerthesaurus.org/sport/synonyms, [Accessed: July 5, 2019].

[55] Shapira, Y., Shapira, B. and Shabtai, A., 2013. Content-based data leakage detection using extended fingerprinting. *2013.*

[56] Costante, E., Fauri, D., Etalle, S., Den Hartog, J. and Zannone, N., 2016, May. A hybrid framework for data loss prevention and detection. In *2016 IEEE Security and Privacy Workshops (SPW)* (pp. 324-333). IEEE, 2016.

[57] Li, Y., McLean, D., Bandar, Z.A. and Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge & Data Engineering*, (8), pp.1138-1150, 2006.

[58] Liu, H. and Wang, P., 2014. Assessing text semantic similarity using ontology. *JSW*, *9*(2), pp.490-497, 2014.

[59] Chahal, P., Singh, M. and Kumar, S., 2015. Relation based Measuring of Semantic Similarity for Web Documents. *International Journal of Computer Applications*, *119* (7), 2015.

[60] Alhindi, H., Traore, I., Woungang, I., “Preventing Data Loss by Harnessing semantic similarity and relevance”, to be submitted to International Conference on Wireless, Intelligent, and Distributed Environment for Communication (WIDECOM 2020), to be held May 06-08, 2020, Toronto, Canada.

[61] Wüchner T, Pretschner A., “Data loss prevention based on data-driven usage control”. In 2012 IEEE 23rd International Symposium on Software Reliability Engineering, pp. 151-160. IEEE, 2012.

[62] Squicciarini, A., Sundareswaran, S. and Lin, D., 2010, “Preventing information leakage from indexing in the cloud.” In *2010 IEEE 3rd International Conference on Cloud Computing* pp. 188-195. IEEE, 2010.

[63] Griffin JL., Jaeger T., Perez R., Sailer R., Van Doorn L., Cáceres R. “Trusted virtual domains: Toward secure distributed services”. In Proceedings of the 1st IEEE Workshop on Hot Topics in System Dependability (HotDep’05). 2005 Jun 30, pp. 12-17.

[64] Burdonov, I., Kosachev, A. and Iakovenko, P., 2009, “Virtualization-based separation of privilege: working with sensitive data in untrusted environment”. In *Proceedings of the 1st EuroSys Workshop on Virtualization Technology for Dependable Systems.* pp. 1-6. ACM, 2009.

[65] Blanke, W.J., 2011, “Data loss prevention using an ephemeral key”. In *2011 International Conference on High Performance Computing & Simulation*. pp. 412-418. IEEE, 2011.

[66] Clark, D., Hunt, S. and Malacaria, P., 2002. “Quantitative analysis of the leakage of confidential data”. *Electronic Notes in Theoretical Computer Science, 59*(3), pp.238-251, 2002.

[67] Yoshihama, S., Mishina, T. and Matsumoto, T., 2010. “Web-Based Data Leakage Prevention”. In *IWSEC (Short Papers)*, pp. 78-93, 2010.

[68] Borders, K. and Prakash, A., 2009, “Quantifying information leaks in outbound web traffic”. In *2009 30th IEEE Symposium on Security and Privacy.* pp. 129-140. IEEE, 2009.

[69] Boehmer, W., 2010, “Analyzing human behavior using case-based reasoning with the help of forensic questions”. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications.* pp. 1189-1194. IEEE, 2010.

[70] Kantor, A., Antebi, L., Kirsch, Y. and Bialik, U., Check Point Software Tech Ltd, 2012. “*Methods for document-to-template matching for data-leak prevention”*. U.S. Patent 8,254,698.

[71] Kale, S.A. and Kulkarni, S.V., 2012. “Data leakage detection”. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(9), pp.668-678, 2012.

[72] Shu, X. and Yao, D.D., 2012, September. “Data leak detection as a service”. In *International Conference on Security and Privacy in Communication Systems.* pp. 222-240. Springer, Berlin, Heidelberg, 2012.

[73] Marecki, J., Srivatsa, M. and Varakantham, P., “A decision theoretic approach to data leakage prevention”. In *2010 IEEE Second International Conference on Social Computing*, pp. 776-784. IEEE, 2010.

[74] Gomez-Hidalgo, J.M., Martin-Abreu, J.M., Nieves, J., Santos, I., Brezo, F. and Bringas, P.G., 2010, “Data leak prevention through named entity recognition”. In *2010 IEEE Second International Conference on Social Computing*, pp. 1129-1134. IEEE, 2010.

[75] Mogull, R. and Securosis, L.L.C., 2007. “Understanding and selecting a data loss prevention solution”. *Technical report, SANS Institute*, p.27, 2007.

[76] Sokolova, M., El Emam, K., Rose, S., Chowdhury, S., Neri, E., Jonker, E. and Peyton, L., 2009, September. "Personal health information leak prevention in heterogeneous texts". In *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, pp. 58-69, 2009

[77] Carvalho, V.R., Balasubramanyan, R. and Cohen, W.W., 2009, July. "Information leaks and suggestions: A case study using mozilla thunder bird". In CEAS 2009-Sixth Conference on Email and Anti-Spam, *2009*.

[78] Kaur, K., Gupta, I. and Singh, A.K., 2018. "Data leakage prevention: e-mail protection via gateway". In *Journal of Physics: Conference Series* (Vol. 933, No. 1, p. 012013). IOP Publishing, 2018.

[79] Ling, H., Chen, Z., Yu, C. and Cheng, Z., SonicWALL Inc, 2018. "Unified source user checking of TCP data packets for network data leakage prevention". U.S. Patent Application 10/015,145, 2018.